DOCUMENT RESUME

ED 131 055 95 SP 010 615

AUTHOR	Collet, LeVerne S.; And Others
TITLE	Formative and Summative Evaluation of the FEHR-Practicum Training Module. Final Report.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO	BR-2-0035
PUB DATE	Jun 75
GRANT	OEG-0-72-0529
NOTE	303p.
EDRS PRICE	MF-$0.83 HC-$16.73 Plus Postage.
DESCRIPTORS	*Computer Assisted Instruction; Computer Programs; Educational Research; Educational Technology; Formative Evaluation; Game Theory; *Practicums; *Program Evaluation; *Research Tools; Simulated Environment; *Simulators; Summative Evaluation; Technological Advancement
IDENTIFIERS	FEHR Practicum; Formative Evaluation Heuristic Research Practicum

ABSTRACT

The purpose of this project was to evaluate the FEHR (formative evaluation and heuristic research) Practicum training model, a computerized simulation providing practical experience in decision-oriented educational research and evaluation. The report is organized into five chapters. Chapter one contains an introductory discussion of the needs and purposes served, a description of a practicum session, and specifications for each of the physical components of the system. Chapter two contains a description of the computer program which generates FEHR Practicum Data, and presents evidence of its portability and adaptability. Chapter three describes the evolution of the present set of simulation problems and provides evidence of the internal validity of each problem. Chapter four presents the results of the empirical evaluation of the FEHR Practicum system in a variety of instructional roles. The fifth and final chapter provides a summary of the evidence regarding the system's effectiveness and discusses the implications for its dissemination and use. The findings provide evidence that, correctly used, the FEHR system can be useful in teaching research/evaluation skills. Its flexible form has proven quite effective for creative instructors who are willing to adapt their methods to the problem-solving mode that is most compatible with the FEHR system. The definitive finding of the evaluation was that FEHR projects are most effective when they are an integral part of a training curriculum teaching research/evaluation techniques and principles in a problem-solving discovery mode. (MB)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED131055

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: SP | IR
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Final Report

Project No. 2-0035
Grant No. OEG-0-72-0529

LeVerne S. Collet
Director & Principal Investigator

and

Robert Parnes
Systems Programmer

With

Elizabeth Anderson
D. Lynne Feagans
Robert Fine
Mary Jane Harlow
Sr. Eileen Rice
Peter Roeper
Nancy Shiffler
Barbara Simon
David Vernon

Recommend approval as final Report
10/15/76

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The University of Michigan
School of Education
Corner East and South University Avenues
Ann Arbor, Michigan 48104

FORMATIVE AND SUMMATIVE EVALUATION
# of the
FEHR-PRACTICUM TRAINING MODULE

June 1975

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development

SP010 615

Final Report

Project No. 2-0035
Grant No. OEG-0-72-0529

FORMATIVE AND SUMMATIVE EVALUATION of the FEHR-PRACTICUM TRAINING MODULE

LeVerne S. Collet
Director & Principal Investigator

and

Robert Parnes
Systems Programmer

With

Elizabeth Anderson
D. Lynne Feagans
Robert Fine
Mary Jane Harlow
Sr. Eileen Rice
Peter Roeper
Nancy Shiffler
Barbara Simon
David Vernon

The University of Michigan
School of Education
Corner East and South University Avenues
Ann Arbor, Michigan 48104

June 1975

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development

ACKNOWLEDGEMENTS

The FEHR-PRACTICUM staff wishes to acknowledge the assistance of Dr. Melvin Semmel and Mr. Jerry Olsen of Indiana University who served as consultants to the project in the early stages. We are most grateful for their assistance and their many valuable suggestions for improvement.

TABLE OF CONTENTS

CHAPTER 1

PAGE

THE FEHR-PRACTICUM SYSTEM....................................1

SECTION I. INTRODUCTION.......................................1
Need..2
Instructional Role..2
Practicum Environment...4

SECTION II. DESCRIPTION OF FEHR-PRACTICUM.....................5
Player's Task...6
Operating Staff...6
Overview of the Game..7
Game Components..11
Experiences Provided by the Game.............................13

SECTION III. FEHR-PRACTICUM MATERIALS........................16
Common Materials...17
Unique Materials...19

SECTION IV. OVERVIEW OF AVAILABLE PROBLEMS...................21
Project PEP: Perceptual Education Problem (RFP001)...........22
Project REMAR: Remedial Arithmetic (RFP002).....24
Project EXTSY: Extended School Year (RFP003)....25
Project Headstart: Early Childhood Education (RFP004)........26
Project TQUEST: Validation of a Teacher Questionnaire (RFP006)................27
Project RMA: Remedial Math for Adults (RFP007)..29
Project BUS: Busing to Achieve Integration (RFP008)..........29

SECTION V. ADAPTING PROBLEM COMPLEXITY TO SUIT CLIENT NEEDS..31

CHAPTER 2

THE FEHR-PRACTICUM COMPUTER PROGRAM.........................33

SECTION I. GENERATION PROCESS................................33
Program Algorithms...37
Steps in Generating Requested Data...........................40

SECTION II. DEVELOPING A PROBLEM PACKET......................48
Phase 1: Operationalizing the Problem..........48
Phase 2: Specifying Program Parameters.........51
Phase 3: Preparing the Problem Packet..........60

CHAPTER 3 PAGE

FORMATIVE EVALUATION AND PROBLEM VALIDATION.................61

SECTION I. FORMATIVE EVALUATION PROCESS....................61

SECTION II. EVOLUTIONARY SYSTEM CHANGES DURING EVALUATION....63
  Introductory Materials..........................63
  The Computer Program............................64
  I.S.T. Units....................................66
  The Message Generator...........................66
  The Information Bank............................66
  RFP Packages (Problem Descriptions)..............67
  Reduced Number of Problems.......................67

SECTION III. ILLUSTRATIVE FEHR PROJECT REPORTS................68
  I. RFP001: Perceptual Education Program (PEP)...69
  II. RFP002: Remedial Arithmetic (REMAR)..........71
  III. RFP003: Extended School Year (ESY)...........84
  IV. RFP004: Headstart (HST).......................90
  V. RFP005: Reading Assessment Problem (READ)....95
  VI. RFP006: Validation of a New Teacher Questionnaire (TQUES)................99
  VII. RFP007: Remedial Math for Adults (RMA)......102
  VIII. RFP008: Busing to Achieve Integration (BUS).104

  INTERNAL VALIDITY OF FEHR-PRACTICUM MODEL.......107

CHAPTER 4

SUMMATIVE EVALUATION.......................................108

  Purpose...108
  General Achievement Objectives...................108
    Objective 1, 2, and 3....................108-109
  General Attitude Objectives......................109
    Objective 4, 5, 6, and 7.......................109
  Summary Objective................................109
    Objective 8....................................109
  Critical Comparisons.............................109
  Contrast 1, 2, 3, and 4....................109-110
  Organization.....................................111

SECTION I. NARRATIVE DESCRIPTION..........................111
  Subjects...111
  Class Settings and Instructional Objectives.....113
  Instrumentation..................................121
  First Examination: A Review of Basic Knowledge..123
  Final Examination: C655, Fall 1972..............124
  FEHR-PRACTICUM Product Rating Sheet.............125

PAGE

Self Assessment of Research and Evaluation Skills (SARES)................................131
Goal Assessment Questionnaire (GAQ)...........132
ORS Questionnaire...........................137

SECTION II. EMPIRICAL EVIDENCE...........................140
Study One: Experimental Evaluation...........142
  Method.....................................143
  Winter Term: Education C656................147
  Results....................................156
  Summary of Experimental Evaluation Results..170

Study Two: Field Trials......................171
  Method.....................................171
  Factor One: Exposure Level.................174
  Factor Two: Problem Content Areas..........196
  Factor Three: Type of Class................211
  Factor Four: Integration...................220

CHAPTER 5

SUMMARY AND CONCLUSIONS..................................228
Satisfaction of the Contract.............................228
Accomplishment of Objectives.............................229
  Objective 1..230
    Experimental Study Results...........................230
    Field Trial Results..................................231
    Amount of Exposure...................................232
    Conclusions..233
  Objective 2..234
    Amount of Exposure...................................235
    Conclusion...235
  Objective 3..236
    Conclusions..238
  Objective 4..238
    Conclusions..240
  Objective 5..240
    Conclusion...241
  Objective 6 and 7......................................241
    Conclusions..241
Problem Content..241
  Conclusion...242
Implications for Dissemination and Use...................243
  Some Philosophical Considerations......................244
  Some Difficulties: Need for Further Development........249

References...251

Appendices PAGE

3A. Budget to Project Report RFP002................252
4A. C655 Final Examination, Fall 1972..............253
4B. Scoring Guide for Product Rating Sheet.........257
4C. SARES Questionnaire............................278
4D. Laboratory Exercises...........................282

LIST OF TABLES

Tables Page

3.1 Mean Scores for Use in Weighting Procedure 88
3.2 Transformed Weighted Scores Used to Ascertain Ranking 89
4.1 Analysis of Covariance of FT Scores Using E1 as Covariate 157
4.2 Profile Analysis of SARES Scores at Time 2 159
4.3 Analysis of Covariance of MC with E1 as Covariate 160
4.4 Analysis of Covariance of MI with E1 as Covariate 161
4.5 Analysis of Covariance of MR with E1 as Covariate 162
4.6 Profile Analysis of SARES Scores for Groups Pooled over Times 2 and 3 164
4.7 Univariate Analysis of Variance of MC 165
4.8 Univariate Analysis of Variance of MI 166
4.9 Univariate Analysis of Variance of MR 167
4.10 Results of the Multivariate Comparisons of Product Ratings for Original C655-C656 and New C656 Subject Groupings 168
4.11 Results of the Multivariate Analysis of Variance of Product Ratings for Pooled Groups 169
4.12 Results of the Multivariate Analysis of Variance of Proposal Ratings Stratified by Exposure Level 178
4.13 Analysis of Covariance of IP Ratings Stratified by Exposure Level (Control Groups Omitted) 179
4.14 Analysis of Covariance of CF Ratings Stratified by Exposure Level (Control Groups Omitted) 180
4.15 Analysis of Covariance of M Ratings Stratified by Exposure Level (Control Groups Omitted) 181
4.16 Analysis of Covariance of GE Ratings Stratified by Exposure Level (Control Groups Omitted) 182
4.17 Analysis of Covariance of the PC Ratings Stratified by Exposure Level (Control Groups Omitted) 183
4.18 Analysis of Covariance of the FC Ratings Stratified by Exposure Level (Control Groups Omitted) 184
4.19 Results of the Multivariate Analysis of Variance of the Goal Attainment Ratings Stratified by Exposure Level (Control Groups Excluded) 186
4.20 Analysis of Covariance of the Attainment Rating Averaged over Goals and Attainment by Exposure Level (Control Groups Omitted) 187
4.21 Results of the Multivariate Analysis of Variance of the Interest Ratings Stratified by Exposure Level. 188
4.22 Analysis of Covariance of the ICR Ratings Stratified by Exposure Level (Control Groups Omitted) 190
4.23 Analysis of Covariance of the IPE Ratings Stratified by Exposure Level (Control Groups Omitted) 191
4.24 Analysis of Covariance of the IPW Ratings Stratified by Exposure Level (Control Groups Omitted) 192
4.25 Analysis of Covariance of the IRP Ratings Stratified by Exposure Level (Control Groups Omitted) 193
4.26 Analysis of Covariance of the Mean of all Interest Ratings. 194

Tables contd.

		Page
4.27	Results of the Multivariate Analysis of Variance of the Difficulty Ratings Stratified by Exposure Level	195
4.28	Analysis of Covariance of the DCR Ratings Stratified by Exposure Level (Control Groups Omitted)	197
4.29	Analysis of Covariance of the DPE Ratings Stratified by Exposure Level (Control Groups Omitted)	198
4.30	Analysis of Covariance of the DPW Ratings Stratified by Exposure Level (Control Groups Omitted)	199
4.31	Analysis of Covariance of the DRP Ratings Stratified by Exposure Level (Control Groups Omitted)	200
4.32	Analysis of Covariance of the Mean of All Difficulty Ratings	201
4.33	Analysis of Variance of the Composite Proposal Ratings Stratified by Problem Content	204
4.34	Analysis of Covariance of the Composite Proposal Ratings Stratified by Problem Content and Covaried on Integration, Exposure, and Class Type	206
4.35	Analysis of Variance of the Composite Final Report Ratings Stratified by Problem Content	207
4.36	Analysis of Covariance of Composite Final Report Ratings Stratified by Problem Content and Covaried on Integration, Exposure, and Class Type	208
4.37	Analysis of Variance of Composite Interest Ratings Stratified by Problem Content	209
4.38	Analysis of Covariance of Composite Interest Ratings Stratified by Problem Content and Covaried on Exposure and Class Type	209
4.39	Analysis of Variance of the Composite Difficulty Ratings Stratified by Problem Content	210
4.40	Analysis of Covariance of the Composite Difficulty Ratings Stratified by Problem Content and Covaried on Exposure and Class Type	210
4.41	Comparison of the Ordered Problem Means from the Original ANOVA with the Adjusted Means from the ANCOVA for Four Variables.	212
4.42	Results of the Multivariate Analysis of Variance of the Observed Proposal Ratings Stratified by Class Type	215
4.43	Analysis of Covariance of the Composite Final Report Ratings Stratified by Class Type and Covaried on Exposure and Integration	217
4.44	Analysis of Covariance of the Composite Interest Ratings Stratified by Class Type and Covaried on Exposure and Integration	218
4.45	Analysis of Covariance of the Composite Difficulty Ratings Stratified by Class Type and Covaried on Exposure and Integration	219
4.46	Analysis of Covariance of the Composite Proposal Ratings Stratified by Integration Level and Covaried on Exposure and Class Type	223

Tables contd.

Page

4.47 Analysis of Covariance of the Composite Final Report Ratings Stratified by Integration Level and Covaried on Exposure and Class Type 224

4.48 Analysis of Covariance of the Overall Means of Interest Ratings Stratified by Integration Level and Covaried on Exposure and Class Type 225

4.49 Analysis of Covariance of the Overall Means of Difficulty Ratings Stratified by Integration Level and Covaried on Exposure and Class Type 226

LIST OF FIGURES

Figures		Page
1.1	Components of FEHR-PRACTICUM	8
1.2	FEHR-PRACTICUM Materials Categorized by Access and Generality	18
2.1	Function of the Keys (or Switches) Entered in Step One of a Request	41
2.2	Summary of the Internal or Construct Variables Contained in the INT File	54
4.1	Summary Description of Participating Classes	122
4.2	Summary of Instruments used in the Summative Evaluation	123
4.3	Matrix of Correlations Among Mean Competency, Interest, and Importance Ratings Collected One Week Apart	133
4.4	Pre-post Correlations for ORS Subscales	139
4.5	Summary of Instrumentation	141
4.6	Distribution of Subjects in the Experimental Evaluation	144
4.7	Schematic Representation of the Experimental Design	148
4.8	Variables Measured at Each Observation Time	150
4.9	Data Classification for the Study Two Field Evaluation	173
4.10	Diagramatic Summary of the Data Matrix Stratified by Exposure Level	176
4.11	Diagramatic Summary of the Data Matrix Stratified by Problem Content	203
4.12	Diagramatic Summary of the Data Matrix Stratified by Type of Class	214
4.13	Diagramatic Summary of the Data Matrix Stratified by Integration Level	222

CHAPTER 1

THE FEHR-PRACTICUM SYSTEM

This report is organized in five chapters. Chapter one contains an introductory discussion of the needs and purposes served, a description of a practicum session, and detailed specifications for each of the physical components of the system. Chapter two contains a description of the computer program which generates FEHR-PRACTICUM data, and presents evidence of its portability and adaptability. Chapter three describes the evolution of the present set of simulation problems and provides evidence of the internal validity of each problem. Chapter four presents the results of the empirical evaluation of the FEHR-PRACTICUM system in a variety of instructional roles. The fifth and final chapter provides a summary of the evidence regarding the system's effectiveness, and discusses the implications for its dissemination and use.

SECTION I. INTRODUCTION

The purpose of this project was the formative and summative evaluation of FEHR-PRACTICUM, which was developed under contract number OEC-0-70-4773(520) with the U. S. Office of Education during 1970 and 1971. FEHR-PRACTICUM is a computerized simulation which provides practical experience in decision-oriented educational research and evaluation. It is intended as a pedagogical tool to facilitate instruction in such program-evaluation tasks as defining the problem, operationalizing objectives, designing valid field studies, budgeting, writing proposals, analyzing data, and interpreting outcomes with respect to an impending decision. The acronym FEHR (pronounced "fair") stands for formative evaluation and heuristic research. Formative evaluation refers to an assessment during the development of a program which performs the functions of feedback, diagnosis, and guidance. Heuristic research is meant to suggest a decision-oriented process that seeks practical solutions to educational problems. The name FEHR-PRACTICUM was intended to emphasize our focus on a practical problem-solving experience which features the use of research/evaluation technology in making decisions about educational programs.

Need

In late 1969, education entered an era in which its sources of revenue began to dry up while its costs continued to climb at an accelerating rate. The combination produced an inexorable demand for educators to provide evidence that their programs were, in fact, producing the results for which they were intended. Simultaneously, educators themselves, faced with austerity budgets, began to clamor for information which would help them decide which programs were most effective and efficient and, alternatively, which could be most easily sacrificed. Many were surprised to discover that personnel who could supply relevant convincing information were largely unavailable -- despite the intensive national research training effort of the sixties.

The reasons for this apparent failure are discussed in a comprehensive report by the Phi Delta Kappa National Study Committee on Evaluation (Stufflebeam, et al., 1971, pp. 302-307). Collecting valid information for this kind of educational accountability, they say, required personnel who are skilled in adapting and integrating the ideas and methods of classical educational/psychological research, economics, political science, administration, decision theory, and general systems theory to meet the specific needs of an impending educational decision. Persons with these skills are hard to find -- even among the graduates of doctoral programs in educational research/evaluation at our most prestigious institutions. Although they identify certain concepts and techniques which need further development, the PDK Committee points out (pp. 307-308) that most major universities currently offer courses which could develop most of the required conceptual skills. What is missing, they say, is a carefully planned sequence of apprenticeship or practicum experiences which can be completely integrated with the instructional activities of the regular curriculum.

Instructional Role

The traditional apprenticeship or practicum experience is unsuited to the training task described above for two reasons: (1) it is usually too far removed from the classroom in both time and distance to permit either direct application of the principles studied or planned reinforcement activities, and (2) the sequency of activities dictated by the needs of the project seldom coincide with the instructional objectives of the training program. However, FEHR-PRACTICUM permits students or practicing professionals who wish to upgrade their skills to get practical experience in a variety of realistic decision-oriented field studies ranging from the validation of a questionnaire for student evaluation of teaching at the college level to the assessment of an elementary reading program or the evaluation of a Headstart project. Each of the above examples is based on a FEHR-PRACTICUM problem. There are eight major problems available, each set in a different content area and involving subjects at a different educational level.

It is important to understand at the outset that FEHR-PRACTICUM is not intended to provide instruction in research/evaluation techniques. Rather, it provides an opportunity to apply theoretical principles to practical educational problems, to practice and develop research/evaluation skills in a complex environment which requires constant extension, generalization, and adaptation of those principles. Pedagogically, FEHR-PRACTICUM is a manageable field experience which is always accessible. It provides a safe vehicle for practicing complicated research strategies, and it provides immediate feedback on the effects of long-term treatments. When carefully articulated with an appropriate training program, the practicum can provide a thread of continuity about which disparate ideas coalesce, thus promoting integration and synthesis.

However, FEHR-PRACTICUM, like other field experiences is not particularly fruitful in isolation. If the practicum is not accompanied by planned instruction, it is imperative that the player-trainees have access to expert consultants and/or ample reference materials assigned for independent use. A discussion of the instructional implications of integrating FEHR-PRACTICUM into an existing training program is provided in a subsequent section.

Practicum Environment

An important advantage of a FEHR-PRACTICUM is that it allows participants to try out new approaches to planning, budgeting, and evaluation without subjecting real student-subjects to the uncertainties of experimental conditions. Instead, the subjects for FEHR-PRACTICUM experiments are drawn from the simulated school system of Fair City, which is located in the mythical state of Utopia, U. S. A. The instructional effectiveness of the practicum is directly related to the quality and depth of this simulated environment.

Fair City. The excitement and challenge of real-life research derive from its potential for improving the educational experience of human beings. Early in the development of FEHR-PRACTICUM, it was discovered that a simulation's capacity to provide this motivating human dimension is heavily dependent upon the degree of contextual detail. Consequently, a great deal of effort was spent in constructing a "community" of sufficient complexity to provide the environment for a variety of educational problems. Participants in the game are given a comprehensive description of the community in the form of an illustrated publication produced by Fair City's "Chamber of Commerce". A copy of this publication appears in a subsequent section of this manual.

Fair City and the state of Utopia are composites of several real cities and states which were carefully chosen to represent the various geo-political sections of the United States. Like most American cities, Fair City has recently experienced a period of rapid growth, with an especially large increase in the black population. In the last dozen years, it has changed from a sleepy town of some 40,000 souls to a bustling city of more than 120,000. The immigrant blacks, being poor, usually settled in the old central area of the city, a region crowded with decaying tenements. The largely white suburbs, on the other hand, are replete with manicured lawns and back yard swimming pools. Many of the educational problems with which FEHR-PRACTICUM players will be concerned derive from these social conditions.

The format of the FEHR-PRACTICUM problems was specifically designed for flexibility. Basically, this was accomplished by providing a checklist of optional assignments which allow the Game Manager to adjust both the scope of the practicum as a whole and the complexity of each of the tasks involved. Guidelines for choosing the options best suited to the instructional purposes of a particular practicum session are provided in section III of this manual.

The extremely flexible structure of the FEHR-PRACTICUM is illustrated by its use as the core experience in each of the following training activities:

1. A twelve-session Saturday morning extension course (workshop) designed to acquaint educational administrators with the basic principles of empirical program evaluation. The course emphasized problem conceptualization skills and the ability to communicate with statistical consultants. There was little formal instruction in the course: the course content was transmitted primarily through intensive consultation during the problem solving process.
2. A one-semester laboratory practicum designed to acquaint first-year graduate students in Special Education with the strengths and weaknesses of various standardized tests commonly used to diagnose learning disabilities, the principles of differential diagnosis, and the basic ideas of research design and statistical analysis.
3. A two-semester sequence of research design and data analysis courses required of all Ph.D. students in education. Most of these students had no previous research experience, and many had previously established negative attitudes towards mathematics and were openly anxious at the prospect of learning statistics.

SECTION II: DESCRIPTION OF FEHR-PRACTICUM

In this section, a general description of the overall game is given, followed by a more detailed explanation of the various game components. We will describe a complete and comprehensive practicum session to illustrate all aspects of the FEHR model. However, the reader should be aware that in practice any coherent subset of the tasks may be assigned. Throughout this description it is assumed that players have been organized into research teams. Although it is possible to play FEHR-PRACTICUM as individuals, or for an individual to play by himself, experience has shown that a richer, more meaningful experience is obtained with the give and take of group decision-making. Consequently, a session usually consists of two or more teams, with each team consisting of from two to five members.

Player's Task

At the beginning of a FEHR-PRACTICUM play each participant is given a Player's Handbook consisting of a brief narrative description of the game, an illustrated description of Fair City, a set of programmed instructions for playing the game, an RFP (Request for Proposals) package containing detailed information about the team's task on this play of the game, and copies of the various request forms used to play the game.

In general, the RFP package provides a verbal statement of an educational problem, identifies the set of (simulated) students involved in the problem, describes several alternative treatments (educational programs) designed to attain that objective, and lists the tests and other instruments which may be used to gather information. The players are then asked to determine empirically which of the alternative programs can best meet the stated objective.

The teams are free to attack their problem in any way they wish, efficient or otherwise. But, throughout the game their actions are subject to the same rewards and frustrations that could be expected in real life. Many of these derive from the fact that research costs money. For example, a team may have too small a budget to permit them to study all the variables they think are important, or they could even have their budget cut by the School Board.

Operating Staff

In addition to the participating teams, at least two staff members are required to operate FEHR-PRACTICUM: a game manager and one or more research consultants. The part played by each staff member is explained below.

The game manager acts as liaison between players and the game components which simulate the educational system. In the field evaluations, the game manager was usually a graduate student, familiar with computers, who had been given two or three hours training in using the physical components of the practicum. The game manager is responsible for collecting monies charged for information (e.g., giving a test to 100 students), and keeping the financial records.

A research consultant serves the same functions as he/she would in real life. Whenever a team is uncertain about research methodology, it may hire a consultant to help. At the beginning of each session, the game manager provides a vita on each available consultant to help the teams decide which person to hire for any one task. A consultant may be hired at any time during the game, providing one is available -- at any point in time it is possible for all the consultants to be engaged by other teams. The cost of consultant service will vary according to the qualifications of the person concerned.

In our evaluation trials, FEHR-PRACTICUM was frequently used in conjunction with a course on research methods. In this arrangement a consultant is unnecessary: the instructor and the contents of the course *per se* perform that function.

Overview of the Game

In FEHR-PRACTICUM each team is hired to "solve" a research/evaluation problem. Throughout the problem-solving process, the teams must collect information about past research in the area and about the behavior of the research subjects (students and teachers). In FEHR-PRACTICUM, synopses of previous research are printed in the Information Bank, a kind of simulated library, and both the research environment (Fair City) and the behavior of the research subjects are simulated by a computer program. Therefore, the teams *cannot visit* the research site in the usual manner. Instead, they must collect their information via the game manager, who might be thought of as a special information line which connects the teams to the simulated school system. This characteristic of FEHR-PRACTICUM is illustrated in figure 1.1.

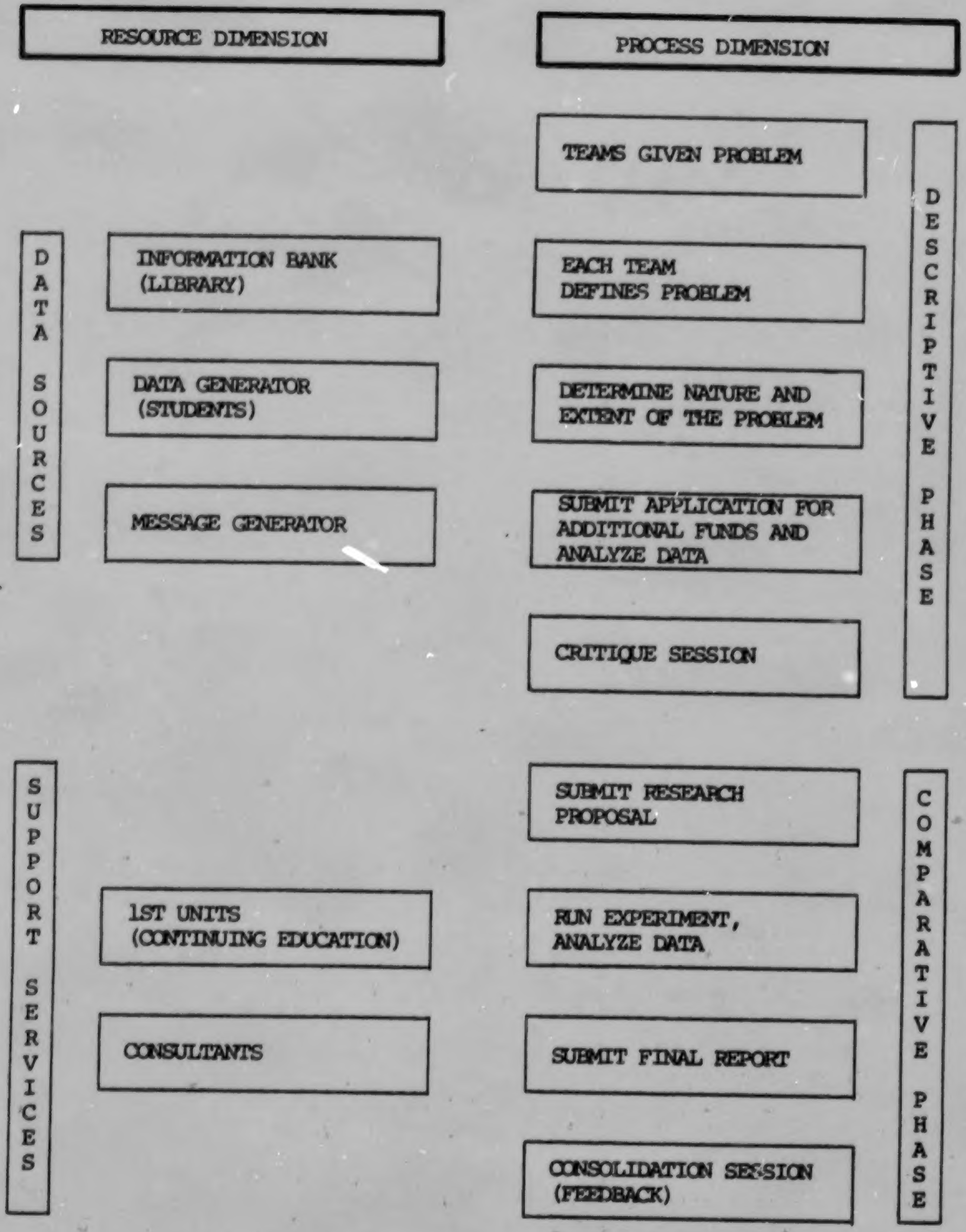

RESOURCE DIMENSION
PROCESS DIMENSION
TEAMS GIVEN PROBLEM
DATA SOURCES
INFORMATION BANK (LIBRARY)
EACH TEAM DEFINES PROBLEM
DATA GENERATOR (STUDENTS)
DETERMINE NATURE AND EXTENT OF THE PROBLEM
MESSAGE GENERATOR
SUBMIT APPLICATION FOR ADDITIONAL FUNDS AND ANALYZE DATA
CRITIQUE SESSION
DESCRIPTIVE PHASE
SUPPORT SERVICES
SUBMIT RESEARCH PROPOSAL
1ST UNITS (CONTINUING EDUCATION)
RUN EXPERIMENT, ANALYZE DATA
CONSULTANTS
SUBMIT FINAL REPORT
CONSOLIDATION SESSION (FEEDBACK)
COMPARATIVE PHASE

Figure 1.1. Components of FEHR-PRACTICUM

FEHR-PRACTICUM is best viewed in two dimensions, the resource dimension and the process dimension, as illustrated in figure 1.1. On the right the process by which a team "solves" an evaluation problem is defined. On the left are the physical components from which the players obtain the information required for their problem-solving activities. Typically, the solution process can be divided into two parts: which we shall refer to as the descriptive phase and the comparative phase.

In the descriptive phase, each team is concerned with obtaining an adequate definition of the problem -- its nature, severity, and extent -- and in determining what other people (i.e., past researchers) have done to remediate the problem. To accomplish this task, players must review the past research in the field, relate the research findings to their knowledge of the community in which the problem is set (Fair City), and conduct surveys (via the Data Generator) using appropriate tests to determine how many students are affected and how severe the problem is. After each team has reported its findings (the descriptive report), all teams working on the problem meet together with the game manager and the consultant(s) in a critique session at which each team's report is critically examined.

During the comparative phase, each team is required to design and conduct an "experiment" to compare the effectiveness of the available treatments with students of various characteristics. The teams then analyze the results, and decide which treatment the schools should use with each type of student. Each team's decision is submitted to the game manager who "operates" the system with that decision in the computer simulator. The computer has the capacity to try one treatment with a student, then set him back where he started and try another treatment. It is therefore possible to compute each team's "decision effectiveness index," which is the ratio of the total growth (learning) obtained under the team's decision to the total growth possible if each student were assigned to the treatment which maximized his growth. In addition, the computer prints, for each available treatment, a set of summary statistics which describe the characteristics of the students whose growth was maximized by that treatment. At the end of the game, the game manager, consultant(s), and players meet together in a consolidation session at which the decision results of each team are critically evaluated and the methodological implications discussed.

Players can best comprehend the nature and scope of the research projects possible in the FEHR-PRACTICUM system by imagining that a real school system exists at the other end of the information line, and that the computer program is a "research assistant" who will do exactly what they ask -- no more, no less. Although it is not able to converse with players, the program can perform the following tasks:

(1) Search the school files and return information such as the grade and past or present achievement scores for an individual student, or for all students in a particular school or class.

(2) Administer tests, attitude scales, or questionnaires to individuals or to a group of students and return the resulting scores. However, in any one FEHR-PRACTICUM problem, the only tests which can be administered are those listed in the variable catalog which is provided at the beginning of the game.

(3) Find and print out the names (ID numbers) of subjects who have patterns of variable scores of a pre-specified type. For example, it could print out the ID's of all students in grade 7 who are male and had IQ scores less than 100.

(4) Administer any specified treatment (educational program) to students identified by individual ID's or to groups of students identified by school, class or a pre-specified pattern of variable scores. Since tests can be administered at any time, they can be used to determine the effects of a treatment over time.

In FEHR-PRACTICUM research, as in real life, the type of research design chosen is frequently dependent on the amount of money available for research. At the beginning of the game, each team is given a research grant. Throughout the game, each test administered and each treatment applied has a cost attached. Teams pay for these services using a special FEHR-PRACTICUM checkbook, which is set up to help them keep track of the monies spent. Thus, one of a team's major tasks is to plan its research so as to ensure that it obtains sufficient information to permit a valid decision without exceeding its grant funds.

FEHR-PRACTICUM is a _game_ in that several teams normally attack the same problem and compete for the "best" solution. However, the competition is parallel rather than direct, since the actions a team takes cannot affect another team's solution in any way. It should be pointed out that there is no "right" experiment to perform and no predetermined "correct" decision. In addition, a team need _not_ decide to use the same treatment for all subjects; it is entirely reasonable to recommend that the schools use different treatments for students with differing characteristics. Whatever decision is made, it will affect students' scores on various achievement tests, attitude scales and the like. Since several teams attack the same problem, it is possible to assess the relative merits of the teams' research procedures by comparing the results obtained by "operating" their decisions in the simulated system. This capacity for feedback on the quality of a researcher's work is considered one of the most valuable aspects of the FEHR-PRACTICUM model.

Game Components

The physical components of the game are of two types: those that are sources of information, and those that teach players how to use the information. The Information Bank, Data Generator and Message Generator supply information, the In-Service Training (IST) Units and the consultants -- the human component -- help the players use the information to "solve" their problem. In the sections below, each of these is described in more detail.

(1) _Information Bank_. The Information Bank is actually a cross-referenced file. Historical information about the Fair City system, statistical data about the tests used in the problems and summaries of the real world studies referred to in the problem booklets are stored in booklet

form for each problem. Several of these are available at each practicum site.

(2) Data Generator. The Data Generator is a computer program which simulates the behavior of the individual subjects within our educational system. Each subject is described by a unique set of scores for a large number of variables such as sex, age, number of siblings, sibling position, intelligence and various attitude scales and achievement tests. However, an individual's scores for many tests -- especially attitudes and achievement -- will change over time. The direction and rate of these changes represent the "growth" or "learning" of individuals over time. In the FEHR simulation, each of the dependent variables has a unique growth curve for each individual student. In addition, each of the available treatments affects the growth curves of the various dependent variables in a different way.

Within the Data Generator, three types of information gathering processes may be used:

(a) File Search. A file search will retrieve the information which exists in the files of the school system. This may be considered fixed information in that teams will get precisely the same scores for each individual each time they search.

(b) Survey. Data which can be obtained by administering a test at the present time is available through a survey. Since measurement error is involved, a survey will return a slightly different test score for each student each time it is used.

(c) Treatment. The treatment process enables the player to administer any available test to subjects at various points in time. Since players may also control which treatments are given to any individual (or group), this process enables you to assess the effects of various treatments experimentally.

(3) Message Interrupts. It frequently happens that a research project is radically changed by external events which the experimenter cannot anticipate or control. For example, a teacher strike which interrupts an experiment may change pupil attitudes as well as introducing costly delays. Such "acts of God" may be introduced into the FEHR-PRACTICUM game by the message interrupts. At various times during the game a team may be given a message by the game manager. Some of these will be relatively unimportant and require no action by the team. Others, however, may require them to make adjustments in their research plan. For example, a message that the research budget has been cut might necessitate the use of smaller samples. Such messages are intended to provide experience in dealing with the unexpected. Message interrupts are an optional feature, to be used at the discretion of the game manager.

Experiences Provided by the Game.

FEHR-PRACTICUM is intended to provide a wide range of practical experience in educational research and evaluation without the expenses and time commitments involved in real research. The practicum provides players with direct experience in gathering and analyzing empirical data in order to arrive at a practical educational decision, and provides feedback respecting the adequacy of their decisions.

Given the goal of simulating the entire research/evaluation experience, common sense would dictate that the closer the simulation is to reality, the more valuable will be its contribution to practice. Consequently, a conscious attempt is made to provide many of the complex interactions (and frustrations) which are characteristic of field research as opposed to laboratory research. A partial list of the experiences which can be provided appears below. However, the user has the option to emphasize one experience and de-emphasize (or omit) another. Instructors may choose the combination best suited to their needs. FEHR-PRACTICUM has the capacity to provide a practical experience in each of the following areas:

1. Identifying and making explicit the basic or "real" problem. Conceptually, a "real" problem may be defined as the discrepancy between what is happening and what should be happening. However, it is a common occurrence in real-life evaluation work to be presented with a "problem" which is, in fact, a request for an implementation decision about one of a series of alternative solutions. For example, "Should we implement program X?" is a solution masquerading as a problem. Identification of the basic problem facilitates the identification of relevant dependent (or criterion) variables. Note: This is perhaps the most difficult task (conceptually) in the entire practicum. Questions of relative value and whose value system must be dealt with.
2. Stating a problem in operational terms. The practicum provides considerable practice in this since the computer requires all requests for information to be made in terms of the values of particular variables.
3. Preparing a budget and working within its constraints. In all except a few restricted versions of problems, the players are given a finite research grant and must pay for each bit of information they collect. In addition, players must pay themselves a daily salary. Thus, careful planning of expenditures of both time and money is necessary.
4. Developing and following a sampling plan. The average FEHR-PRACTICUM problems contains literally thousands of potential research subjects, each with a wide variety of individual characteristics (sex, intelligence, socio-economic status, etc.). Almost any sampling plan which can be used in real research can be duplicated in the game -- including plans which are invalid because of some type of selection bias.
5. Selecting dependent and independent (moderator) variables which are relevant to a given problem and choosing the instruments (tests) which will be used to measure them.

Although the players cannot devise their own tests, they may choose from a large pool of tests which are made available in the practicum. To help them in assessing the utility of the various tests, players have access, via the Information Bank, to test descriptions of the sort provided by Buros (1965). Depending on the problem area, the game provides scores on from 50 to 160 separate tests. Each test may be used with any subject, and may be administered repeatedly across time.

6. Using survey techniques to identify the important dependent and independent variables in a given educational problem. In the practicum, surveys are frequently required to determine the extent and severity of a problem and/or to clarify relationships among variables.

7. Designing research plans which isolate the effects of specific educational treatments and treatment combinations. The practicum allows players to collect data according to almost any research design which can be used in a real-life situation -- including biased or invalid designs. The possible designs include both the univariate and multivariate forms of latin squares, incomplete blocks, longitudinal studies (panel data), and case studies based on variable scores (rather than verbal descriptions). Because of the capacity to produce longitudinal data, it is possible to stimulate formative evaluation studies involving sequences of treatments and repeated observation periods.

8. Analyzing data collected from complex designs. The capacity to provide such analytic experience is ensured by the complex designs mentioned in (7) above. In addition, the game has a number of built in biases which encourage players to use designs involving multiple criteria (dependent variables). Thus, multivariate analyses are usually appropriate. (Of course, the capacity actually to conduct such analyses depends on the resources of the local computer installation.)

9. Modifying research plans to accommodate unforeseen events in the environment. For example, a teacher strike could modify student attitudes as well as cause an expensive delay in a project. Such simulated events can be used with sophisticated trainees, but are not recommended for beginners.

10. Selecting consultants and preparing plans to optimize their effectiveness. The practicum provides an opportunity for players to explore their own limitations, and to find the conditions under which a consultant is "worth the money".

11. Relating the results of an evaluation to the time at which the evaluation is taken. The game permits program evaluations to be made at different points in time and to compare the results.

12. Working with educational problems in a variety of content areas and at numerous educational levels. The topics of the eight problems available run from the traditional subjects (e.g., mathematics and reading) to the specialized difficulties of handicapped children. The educational levels represented include both pre-school children, and college students.

13. Assessing the quality of research procedures (in the comparative phase) by examining the results obtained from "operating" a decision based on the research results. To aid in this task, the computer supplies two bits of information which are not obtainable in real-life research: the "decision effectiveness index" and a statistical summary of the characteristics of the students best suited to each treatment. These are not intended as absolute indices of quality, but rather as springboards for discussion.

SECTION III. FEHR-PRACTICUM MATERIALS

The FEHR-PRACTICUM materials can be classified on two broad dimensions. The first of these is the access dimension. Where is the material physically located? How and by whom is it normally accessed?

The second dimension concerns the generality of the materials, whether it can be used in all problems (content areas) or not.

The access dimension is sub-divided into four categories. Category 1 contains all the data generator materials. These would normally be accessed at the local computer center or a remote terminal. Category 2 contains materials which would normally be used only by the Game Manager and/or those planning the instructional uses of the practicum. Category 3 contains materials which are shared among players. These would normally be accessed in a laboratory-classroom. Category 4 would contain all the materials normally provided to each player-trainee.

The generality dimension contains two categories. Category 1 (common) contains all materials which can be used with all eight problems, while category 2 (unique) contains materials which change from problem to problem.

The entire set of FEHR-PRACTICUM materials, categorized by access and generality, appears in figure 1.2. In the discussion below, there is additional descriptive information for every component except the Game Manager's Manual which has been described in context.

Common Materials

The main data generator (main computer program) consists of a set of punch cards (or a computer tape) containing the FORTRAN IV source program. However, the main program cannot produce simulated data unless it is combined with one of the data generator problem pockets (see below).

The FEHR-PRACTICUM IST (in service training) units consist of separately-bound, semi-programmed materials which provide detailed instructions for accomplishing specific tasks encountered in the practicum. The five units available are:

I. Assessing Success for Complex Objectives.

II. Criteria for Developing Proposals and Final Reports.

III. Computer Format Statements for FEHR Data.

ACCESS DIMENSION	GENERALITY DIMENSION	
	1. COMMON TO ALL PROBLEMS	2. UNIQUE CONTENTS FOR EACH PROBLEM
1. COMPUTER CENTER	a. Main Data Generator (Main computer program.)	a. Data Generator Problem Packet: unique program parameters for each problem. (Eight separate packets.)
2. GAME MANAGER	a. Game Manager's Manual Sections I & II	a. Game Manager's Manual Sections III & IV
3. LABORATORY OR CLASS-ROOM	a. FEHR-PRACTICUM 1ST Units. (Five separately-bound units.) Note: Some Game Managers may wish to supply a copy of this material to each player. b. References. (Supplied locally.)	a. Information Bank: material separately bound for each problem. (Eight separate Information Banks.)
4. PLAYER MATERIALS	a. Player's Introduction to the FEHR-PRACTICUM game. b. Fair City, U. S. A. c. Player's Instructions for FEHR-PRACTICUM.	a. RFP (Request For Proposals) Document: a specific description of the particular problem to be investigated, separately bound for each problem. (Eight separate RFP documents.)

Figure 1.2. FEHR-PRACTICUM Materials Categorized by Access and Generality.

IV. Sampling the Subjects to be Studied.

V. Using the FEHR Secretary.

The *references* to be supplied locally consist of any research-oriented materials which will assist the player-trainees. A list of suggested titles appears in a subsequent section.

There are three player's materials which are common to all problems. All three of these materials appear in this manual. The first, *Player's Introduction to the FEHR-PRACTICUM Game*, was contained in the first twelve pages of Section II, which you have just read. The second and third booklets, *Fair City, U. S. A.*, and *Player's Instructions for FEHR-PRACTICUM* constitute the second and third chapters of Section II. Wherever necessary, the original materials printed herein have been supplemented by notes and addenda addressed to the Game Manager.

Unique Materials

The total FEHR-PRACTICUM system provides a choice among eight major problems dealing with eight different content areas and involving students at different educational levels. Each problem has its own unique Data Generator Problem Packet, Information Bank, and RFP (Request for Proposals) document. Although the specific contents differ, the format of each of these components is the same for all problems. The usual format for each of the three components is described below.

(1) The *Data* Generator Program Packet combines with the *Main* Data Generator (above) to form the complete computer program necessary to operate a specific problem. When the first order is received from a new user, an intact deck containing both the Main Data Generator and the Data Generator Packet for the standard REMAR (Remedial Arithmetic) problem is shipped. When you are familiar with this problem new ones may be ordered one at a time. Since the *main* sections of the program are common across problems, we routinely ship only the Data Generator Problem Packets to users who already have one or more FEHR problems operational.

(2) The Information Bank contains a brief summary for a number of real-life articles related to the content area of the problem, plus descriptive information for each of the standardized tests available in the problem. These are printed on loose-leaf sheets, one article per page, and arranged alphabetically by author. This format was adopted to permit users to update the information as new relevant research becomes available.

(3) The RFP document for each problem has the same general form. Page 1 identifies the content area and the educational agency which is sponsoring the research. Page 2 provides a general narrative description of the problem and refers the reader to an appendix from which more specific details may be obtained. Pages 3 to 6 contain the Checklist of Tasks to be Performed: a detailed listing of all the tasks involved in a complete practicum. The Game Manager chooses those tasks best suited to the local needs.

The detailed content of each RFP is contained in a set of appendices. These are usually five or six appendices containing the information described below:

Appendix I. Information Bank Material. This is a list (or index) of all the abstracted articles which come along with the simulation. Going through these articles will give the player an overview of the research in the area. If a player is especially interested in the substantive area to be investigated, we suggest that the Information Bank be used to determine which articles should be read in full. In addition, the Information Bank provides normative data (means, standard deviations, reliabilities, etc.) and a description of the test content for all standardized tests listed for that problem. In addition, for many tests a critique is also available in the Information Bank.

Appendix II. Research Population. Appendix II contains a complete list of all subjects available in the problem and explains how to interpret the student ID.

Appendix III. Catalog of Treatments. Appendix III is a list of the treatments which can be administered in the problem, a list of costs of each treatment, and a definition of the time sequences used in the problem.

Appendix IV. Catalog of Variables. Appendix IV lists all the variables available in the problem, the costs of each variable score, and the conditions under which the variable scores (test scores) may (and may not) be obtained.

Appendix V. Committee Report. In nearly all problems there is some preliminary information available such as why particular treatments were chosen and what previous research the school system has done in the area. A concise summary of this information appears in Appendix VI.

SECTION IV. OVERVIEW OF AVAILABLE PROBLEMS

A total of eight problems are available in the FEHR-PRACTICUM system. Once the Game Manager and research consultants have become thoroughly familiar with the FEHR-PRACTICUM system, it is possible to conduct a practicum in which several problems (indeed, all eight) are being operated simultaneously. This has the advantage of permitting player-trainees to choose the area closest to their own substantive interests. Although the problem-solving procedures are similar from problem to problem, we have found that the choice of content area can make a tremendous difference to a trainee's motivation. However, managing FEHR-PRACTICUM is a complex task. We strongly recommend that new users not operate multiple problems until they have had at least two or three sessions practice with a single problem.

During our field validation studies, we tried training Game Managers on several different problems. It was our experience that prospective Game Managers who were supervising problems based in a content area with which they were unfamiliar had great difficulty concentrating on the management tasks *per se*, and consequently learned much more slowly than those supervising problems with which they were familiar. Since it was also found that numerous demonstration file searches, surveys, and field experiments facilitated learning, it was decided to develop training materials based on one standard problem and featuring many practical examples. The REMAR problem was chosen for this purpose because remedial arithmetic is the one content area with which most educators have had some experience. Section III of the Game Manager's Manual contains programmed directions and a number of practical examples for each task in the REMAR problem. After completing this section, most prospective users had little difficulty administering a full-fledged practicum session using the REMAR problem. The extension to other problems is then just a matter of becoming familiar with the content area by going through a complete practicum following the Player's Instructions step by step.

The eight available problems are described below in the order they were developed. The instructional strengths and limitations of each problem are listed under the heading *Special Characteristics*. For sessions in which the specific content of a problem is of substantive interest, we strongly recommend that copies of the primary references listed with the problem be made available.

(1) *Project PEP*: *P*erceptual *E*ducation *P*roblem (RFP001). The term "perceptually handicapped" has been used in recent times to identify a large number of children who have normal intelligence but because of a "perceptual problem" have great trouble in school, particularly in reading and writing. In this problem the players are requested to aid the Board of Education and a committee of teachers in deciding such questions as: Does Fair City need a Perceptual Education Program: Are psychological and socioeconomic variables relevant? Which treatment should be recommended?

__Particulars__: This problem has 161 variables, three treatments (2 experimental, 1 control), and a total research population of 426 students in four grades of one school.

__Special Characteristics__: This problem permits a direct experimental comparison of the two proposed programs with present practice (the control). It features the usual sampling, variables selection, and design difficulties, with emphasis on the last two. One of the major difficulties in PEP is problem definition. The School Board sees the problem as a lack of achievement. How does that relate to "perceptual handicaps"? What is a perceptually handicapped child? What variable scores signify a handicap? The conflict between end-result variables (achievement) and intermediate results (changes in variables which are hypothesized as prerequisite to successful achievement) make this problem particularly useful for practicing problem conceptualization skills. An additional feature is the "built-in" experience with regression toward the mean which is caused by stringent selection criteria.

__Primary References__:

Frostig, Marianne. "Visual Perception in Brain-Injured Children". __American Journal of Orthopsychiatry__, 1963, 33 (4), 665-671.

Johnson, J. & Myklebust, H. R. __Learning Disabilities: Educational Principles and Practices__, Grune & Stratton, New York, 1967.

Kephart, Newell C. (U. of Purdue, Lafayette) "Perceptual-Motor Aspects of Learning Disabilities". __Exceptional Children__, 1964, 31 (4), 201-206.

McCarthy, J. J. & McCarthy, J. F. __Learning Disabilities__, Allyn & Bacon, Boston, 1969.

(2) Project REMAR: Remedial Arithmetic (RFP002). This problem deals with methods of teaching arithmetic computation skills. Because most educators are familiar with this content, this was chosen as the "standard" problem to be used for the first implementation at a new site. The RFP is issued by the Fair City School Board, who are concerned at the growing number of grade seven students who cannot do arithmetic computation well enough to succeed in the regular grade seven curriculum. You are asked to conduct field tests to evaluate the effectiveness of three proposed new remedial arithmetic programs as compared to the current practice. At the conclusion of the project, each team must decide on the basis of their experiment which of the new programs (if any) are to be implemented. (Note: it is entirely possible to recommend different programs for students of differing characteristics.)

Particulars: Variables = 78, Treatments = 4, Total Research Population = 1,906 seventh grade students from seven junior highs, each with several classes.

Special Characteristics: This problem permits direct experimental comparisons. It is fairly heavily oriented toward criterion-referenced tests or sequential mastery tests. Sampling, selection of variables, and design all are involved and can be accomplished fairly easily. Because of the necessity to select only the poor students, this problem provides a rich opportunity to study the effects of statistical regression. In addition, there are some conceptual difficulties with respect to the precise definition of success in terms of variable scores, and some sticky statistical questions centering around the analysis of mastery-test data. Nevertheless, this is perhaps the easiest, most straightforward problem, since the objectives are fairly clearly defined in unambiguous terms.

A unique feature of this problem is the emphasis on the cost-effectiveness aspect of the decision among programs which is introduced by a wide disparity in program costs and a positive correlation between cost and learning.

Primary References: (None)

(3) Project EXTSY: Extended School Year (RFP003). What are the benefits of an extended school year? Fair City, like many other areas, believes there would be economic if not educational benefits from having the school schedule reorganized so that the schools are in operation all year around. The players are to investigate the situation and determine which, if any, schedule has the most advantages for Fair City.

Particulars: Variables = 56, Treatments = 3 (2 experimental, 1 control), Research Population = 12,393 students from 21 elementary schools.

Special Characteristics: This problem permits some experimental comparison, but notice that an entire school must be assigned to any one treatment. This introduces some interesting questions with respect to the appropriate unit of observation and the generalizability of results. EXTSY involves extensive sampling, variable selection, and design problems. One of the difficulties in this problem is the fact that since treatments must be administered to intact schools, no true experimental design is possible. Attitude variables may be more valuable here, but reliability and validity problems, as well as the nominal nature of such data, are problematic. The fundamental question is "which is more important: achievement, cost, or popularity"? Several unique aspects of this problem are introduced by its longitudinal nature (3 years).

Primary References:

Department of Education, New York. "The Impact of a Rescheduled School Year". A special report prepared for the Governor and the Legislature of the State of New York. The University of the State of New York, the State Education Department, Albany, New York, March, 1970, 158 pages.

(4) Project HEADSTART: Early Childhood Education (RFP004). The Fair City School District believes that there are a growing number of children who are entering first grade ill equipped to perform at normal levels. In this problem the players must aid the Board of Education in deciding if a Headstart program should be introduced and which particular program best meets the needs of the Fair City children who require extra attention. The players should not only decide if there is a need but also if the gains made in Headstart are retained after the child has entered the regular public schools.

Particulars: Variables - 78, Treatments = 7, Research Population = 1,822 or all three-year-olds in the city.

Special Characteristics: Although direct empirical comparisons among the 7 treatments are possible, this is too complicated to permit in practice. An additional complication is the fact that not many measures are available for pre-school kids, let along many reliable ones, particularly those which may change as a result of some program. In addition, the problem requires long-term (longitudinal) assessment of changes, and most available tests -- even though they have the same name -- have different norms for different age groups.

Primary References:

Weikert, D. "The Perry Preschool Project in Ypsilanti, Michigan", 1969, OE-37035.

(5) Project READ: Reading Assessment Problem (RFP005). The teachers of the primary grades have become dissatisfied with their present reading program because of the increasing number of students who are falling behind their peers in the development of reading skills. In this problem the players are to aid the teachers and principals of the elementary schools to determine if a new reading program should be instituted in Fair City. One of the questions they will answer in this problem is whether there is one curriculum which can best meet the needs of all Fair City children.

Particulars: Variables = 170, Treatments = 3 (2 experimental, 1 control), Research Population = 2,000.

Special Characteristics: This problem permits direct experimental comparison. The major emphasis here is assessment in terms of multiple behavioral objectives. About half the variables in this problem are criterion-referenced. Consequently, variable selection is an important element in this problem. But perhaps the major feature is the data interpretation task. The multiple successes and failures of students in various programs must somehow be summarized in a conceptually meaningful way to permit program-to-program comparisons, and a subsequent decision among programs.

Primary References:

(Not used in development, but very similar and helpful)

Duffy, G. G. & Sherman, G. B. *Systematic Reading Instruction*, Harper & Row, 1972.

(6) Project TQUEST: Validation of a Teacher Questionnaire (RFP006). The purpose of this project is to validate a questionnaire which "evaluates" teacher performance at the college level. The questionnaires are to be administered to students presently enrolled in college classes. The players are to assist the administration in this project

by comparing the effects of feeding back information from various sub-scales on the new questionnaire. A second questionnaire and several achievement scores are also available to be used in the validation task.

Particulars: Variables = 60, Treatments = 4, (3 types of feedback, 1 no feedback), Research Population = 512 university students from 20 different classes.

Special Characteristics: This problem contains two questionnaires: (1) the old questionnaire, and (2) a new one designed to provide more information. It opens up quite a can of worms -- how does one validate such a thing as a student evaluation of teachers? In this problem, the path suggested is to see how effective the questionnaire is in changing professors' teaching, as measured by the questionnaire. The variables (questionnaire items) are all 5 option attitude items. Individual item responses may then be combined to form various scales which relate to the developers' (i.e., the University Committees) concept of teaching effectiveness. Feeding back to a professor his "scores" on one or more of these scales from last semester should influence his score (on the scale(s) concerned) this semester. Thus it is possible to collect evidence of the construct validity of the questionnaire.

It is important to note that, while we give questionnaires to students, the unit of observation is a teacher (i.e., the class). This introduces a variety of interesting statistical questions which are an important aspect of this problem.

Primary References:

Cronback, L. J. Essentials of Psychological Testing, Chapter 5, "Test Validation", Third Edition, Harper & Row, 1970.

(7) Project RMA: Remedial Math for Adults (RFP007). The open enrollment policy and the wide variety of people who attend community colleges necessitates the provision of additional support services for students and citizens. In this problem the players are requested to evaluate the remedial math course at the Fair City Community College. This course is intended to provide the students taking it with the skills to do college level work. The players will be asked to answer such questions as: Does a remedial program work for adults? For what kind of person is this program least useful? How can the program be made more effective and efficient? Should the course be continued?

Particulars: Variables = 26, Treatments = 1, Research Population = 251.

Special Characteristics: This problem is strictly a post hoc evaluation task, with evaluation made solely on the basis of evidence collected during the semester. The trickiest part of this problem is an operational definition of success. Since the students come from a wide variety of backgrounds, and have very different goals, the meaning of success varies from group to group. The selection of relevant variables can also get immensely complex, since there are several varieties of attitude, aptitude, and achievement measures which the teams might be used in any combination.

Primary References:

Dalke, Richard M. A Case Study of an Individualized Course in Arithmetic at a Community College. Dissertation, University of Michigan, 1971.

(8) Project BUS: Busing to Achieve Integration (RFP008). In response to recent Supreme Court rulings, the Fair City Board of Education has decided to integrate the city's schools through busing. As in most cities, the people

of Fair City have very strong views about busing and there are many complications to consider. In this problem the players are requested to evaluate the effects of busing in the city's elementary schools to determine its advantages and disadvantages and what are the sources of the problems that exist.

Particulars: Variables = 52, Treatments = 1, Research Population = 3,633 pupils in grades 1 and 4 of 21 different schools.

Special Characteristics: This problem is strictly a post hoc evaluation. Teams are called in after the busing decision is made. Although only the first and fourth graders are available to the players, there is nevertheless a large number of subjects from a wide variety of socio-ethnic neighborhoods. Since computer costs prohibit an exhaustive survey, this problem offers a rich environment for practicing sampling skills. There is the fact that any results may be attributable to the new organization in the elementary schools rather than the busing plan itself. There is also extensive direct measurement experience, since it may place more weight on attitude measures as criteria. The relatively low reliability and validity of such measures and the categorical nature will necessitate the construction of broader construct variables through various combinations of questionnaire responses.

Primary References:

Sullivan, Neil U. and Steward, Evelyn A. "Now is the Time: Integration in the Berkeley Schools", Indiana University Press, Bloomington, 1969 (ERIC).

SECTION V. ADAPTING PROBLEM COMPLEXITY TO SUIT CLIENT NEEDS

It should be obvious from the foregoing description that a comprehensive FEHR project demands a high level of research expertise. It is, for example, an implicit assumption of FEHR-PRACTICUM that educational evaluation is a multi-dimensional activity. Each problem contains many dependent variables (e.g., achievement tests, attitude scales and the like), several treatments (alternative educational programs) and a wide variety of independent moderator variables (such as sex, race, or socio-economic status). To use the full multivariate capacity of the simulation, participants should be familiar with the classical literature in educational measurement and research design, and able to use multivariate statistical analyses.

In addition, the comprehensive problems described above required the participants themselves to identify the exact nature of the problem and to specify, in detail, the nature of the solution strategy. The large number of sophisticated problem-definition decisions required in this version of the practicum made it an ideal vehicle for training advanced students who were specializing in educational research evaluation. However, many potential users did not possess the prerequisite skills; it was too complicated to use non-specialists or with students just beginning their research training. In addition, the unstructured version generally took fifteen or twenty three-hour sessions to complete, with almost half the sessions spent in defining the problem. To extend the utility of the game, it was desirable to provide shorter and simpler versions of the FEHR problems.

The complexity of a problem (and consequently the time required to complete a practicum session) can be reduced by either structuring or restricting the problem, or both. A problem may be structured by providing an operational definition of the problem. For example, at one of the evaluation sites the REMAR problem was structured by defining a remedial arithmetic student as any student who scores zero on one or more of the mastery tests of computational skill which were available in that problem.

A problem may be restricted either by limiting the number of dependent and independent variables to be included in a study or by requiring a specific set of variables to be used. In the example above, the teams were required to use the SAT standardized test of computational skill as the dependent variable.

It was mentioned previously that each RFP package contains a section titled Checklist of Tasks to be Performed. Below each of the tasks listed in this section are a variety of optional restrictions and structures. The instructor or game manager can vary the complexity of the total project over a wide range simply by checking off different patterns of assigned tasks. In this way it is theoretically possible to adapt the practicum to suit the needs and abilities of any group of prospective clients.

Encouraging Creativity. Many evaluation specialists (Stake 1967, Tyler 1967, Stufflebeam, et. al., 1971) have stressed the need for creative approaches to evaluation problems. In this view, not only the solutions *per se*, but also the solution methodologies are idiosyncratic to problems at hand. For this reason, it seemed desirable for the teams to have some capacity to define the task for themselves -- even in the shorter versions of a problem. In the field tests, the apparent conflict between the need for structure and the desire to retain sufficient flexibility to permit some team creativity was resolved by requiring that certain variables and operational definitions be included in an evaluation study, and simultaneously encouraging teams to add variables and definitions which they believed would increase the validity of their findings. To keep the number of added elements from usurping a good deal of time, budgetary restrictions were imposed. In general, the research grant allotted to each team included the projected cost of the required tasks plus 15 or 20 percent "discretionary funds" which the teams could use to improve the substantive value of their research.

CHAPTER 2

THE FEHR-PRACTICUM COMPUTER PROGRAM

In chapter one the data generator was described as a "research assistant" which can be used to manipulate the simulated educational system being studied in a particular problem. Through the data generator the research team can retrieve information about any simulated student from the school files. The generator can administer any available educational treatment to any student or group of students, and measure the effects of that treatment by administering tests or questionnaires to the students concerned at any time during the treatment (e.g., pre and post testing is possible). This chapter is divided into two sections. Section I is concerned with the general process by which these data sets are generated. Section II describes the procedure for developing a particular problem packet.

SECTION I. GENERATION PROCESS

There are two general methods by which simulated data type have been obtained by previous investigators: random selection from a comprehensive data bank, and random generation via a probability density function suited to the variable concerned. Neither of these procedures proved entirely satisfactory for producing large multivariate data sets with prescribed interrelationships among variables. Consequently, a compromise data generating procedure was developed.

In FEHR-PRACTICUM, each subject in the population available for a particular problem has associated with him a set of variable values which uniquely describe that individual. To maintain an individual identity the values associated with one individual must be highly correlated from run to run: indeed, some variables (e.g., sex) must yield the same score value every time that individual is accessed. This means in effect that the variables cannot be randomly generated from "scratch" each time the program is run. For these variables, parameter values for each individual must be stored permanently to enable the required consistency to be maintained.

At the same time, there are other variables associated with each individual which simulate learning by changing systematically over time in response to the particular educational treatments that are administered. In addition, we desired the effects of each treatment to be modified by the characteristics of the subject to whom it was administered. Since a problem typically contains thousands of subjects, hundreds of variables and several treatments, it is obvious that it would have been impossible to store all the results in a data bank to be accessed on demand.

The FEHR-PRACTICUM data generator incorporates the strengths of both the data bank and random generation methods. The overall strategy was to construct a small "internal" data base which encapsulated the desired interpersonal and intrapersonal relationships. This data base provides the stability that is needed in order to recover for each individual his own unique pattern of scores each time he is referenced. These score patterns ought to be consistent both across time and across the set of variables available in that problem. For instance, the Otis IQ score for a particular simulated individual should remain relatively constant (within measurement error) from one simulated time period to another. Also, a simulated individual who scores high on the Otis IQ ought to score high on the Stanford-Binet IQ. In addition, any simulated individual with a high IQ score (regardless of the test used) ought, in the absence of other moderating variables, to be doing quite well in school. These consistencies are provided by an internal data base with a prescribed pattern of intercorrelations.

Internal Data Base. The internal data base was constructed by a procedure that can best be described as the reverse of a factor analysis. A set of five independent factor scores was randomly generated for fifty individuals. Fifty-three internal variable scores were then generated for each individual by taking different linear combinations of the five factors. The correlation between pairs of internal variables was controlled by the amount of commonality in the combination rules by which the variables were created. In this way we were able to create fifty prototypical individuals each having fifty-three "true" scores which maintained pre-specified intercorrelations. Each variable was then transformed to normal deviate form, and the resulting 50 x 53 matrix of Z scores were stored as a block. Nine such blocks were created, each having a different pattern of intercorrelations. This enables us to simulate homogeneous intact groups with widely differing characteristics from group to group. Each simulated group is referenced to one of the nine blocks, and each subject within the group is referenced to one of the fifty prototypical individuals within that block. Every time the scores for a particular individual are needed, the same set of internal variables is accessed. These internal Z scores are never seen by the research team: they are used internally to generate the external or raw scores which a team receives.

Generating External Scores. There is a distinction in the social sciences among nominal, ordinal, and metric (interval or ratio) scales of measurement. This distinction was built into the FEHR-PRACTICUM data generator by varying the way an internal variable was translated into an external variable. Each external variable was labeled *a priori* as to scale type, and this was coded into the program. When a variable is requested, the code is used to choose the appropriate procedure for translating the internal Z score into an external variable score. The translation is performed every time the variable is requested according to translation parameters stored in the program.

For all metric variables, an individual's internal Z score is first translated to an external score by the formula

$$\text{Score} = [ZI(r) + ZR\sqrt{1 - r^2}]\,\sigma + \mu$$

where ZI is the individual's internal Z score, ZR is a randomly generated Z score, and r, σ, and μ are stored parameters prescribing the reliability, standard deviation, and mean of the external variable being generated. These three parameters must be stored for each external variable. The expression within the square parentheses yields the Z score of the external variable. The first component of the expression derives from the "true" score and the second is the error of measurement. The relative sizes of the two components is controlled by the reliability parameter. The range of raw score values is controlled by the σ and μ parameters.

Ordinal data are generated using the expression within square parentheses in the score formula. Second, each external Z score is then translated to its decile value in the normal distribution. Third, the decile value is matched to an external score. Thus, for each external variable of the ordinal type a reliability parameter (r), and ten decile values must be stored within the program.

Nominal data are generated in two ways. Wherever the external variable is required to change its distribution across categories in response to a treatment effect, nominal data are generated by exactly the same general procedure as ordinal data. The only difference is that the external scores matched to the deciles need not be arranged in order of size. However, there are certain nominal variables for which it was desirable to maintain a more direct control over the correlation between category membership and the scores on other variables. Sex, number of siblings, and ethnic group are examples of this variable type. For these critical variables, the actual category membership was stored as one of the 53 internal variables.

Multiple Use of Internal Z Scores. A single internal Z score can be used to generate the scores on several different tests, providing they have the same underlying construct. For example, all the IQ scores for one individual are generated from the same internal Z score. Similarly, the scores on many different reading tests can be generated from a second internal Z score. Since these two internal variables have a prescribed "true" correlation, the correlation of external variables will tend to have the same value. The amount of variation in the external correlation from sample to sample is a function of the reliability parameters of the external variables concerned.

In the same way that a single internal Z can be used to generate many external scores, it is also possible to use the vector of 53 internal Z scores which describe one prototypical individual to generate many different external individuals (i.e., research subjects). The external scores of subjects generated from the same internal prototype can be quite different because the error component for each score is generated randomly.

The multiple use rationale is also extended to each block of internal Z scores. We can use the same internal block to represent external groups with radically different score patterns (e.g., a grade seven class and a grade eight class) by supplying a different set of variable parameters.

PROGRAM ALGORITHMS

The purpose of this section is to provide a logical description of the specific algorithms used in generating the simulated scores called for by a user's request. The term request as used here is defined as a set of punch cards which relate to a single operation -- that is, to a file search, a survey, or an "experiment" involving the administration of different educational treatments to groups of simulated subjects. Specific instructions for preparing a request are given on pages 100 to 110 of the Game Manager's Manual and in the Players' Instructions on pages 14 to 19. For convenience, the operation of the various program components are described here in the order that they are called by the request deck.

Throughout this description we shall be using a variety of familiar terms which have had somewhat different meanings within FEHR-PRACTICUM. A list of terms with unique meanings appears below.

> Research Population. The research population of a problem consists of all the subjects which can be used in that problem. A complete listing of all subjects in a given problem appears in each Request For Proposals (RFP).
>
> Subject I.D. Each subject is identified by a seven digit I.D. number consisting of three segments. In the general case - the first two digits identify the unit, the next two digits identify the sub-unit, and the last three digits identify the particular subject within a sub-unit.
>
> Units and Sub-Units. Units and sub-units are a general method of identifying particular groups within a research population.

For example, in the remedial arithmetic problem the unit identifies the school which a subject attends and the sub-unit identifies his classroom within that school. In the busing problem, the unit identifies the school a subject is now attending and the sub-unit identifies his former school. The specific meanings of units and sub-units for a particular problem are defined in the RFP.

Catalog of Variables. Each RFP contains a catalog of variables which lists each variable that can be used in that problem. For each variable, the catalog provides the following information: (1) a three digit index used to identify the variable to the computer, (2) whether or not the variable can be used in: a file search, a survey, or a treatment, and (3) the cost of obtaining one subject's score on that variable.

Observation Time. Within the simulation only a limited number of different observation times (testing times) may be provided. Each problem defines a particular beginning time (e.g., the first day of school) as time zero and a time unit (e.g., one week); within the problem all observations must then be expressed as a two digit number representing the number of time units after time zero (e.g., time 05 would mean five units -- weeks -- after the beginning time).

File Search. A file search retrieves each subject's score on particular tests administered some time in the past. Therefore, a file search will return exactly the same score for a particular individual each time it is used. Since no actual testing is involved, file searches are much cheaper than surveys. However, variables which are identified in the catalog of variables as available on file can also be obtained via a file search if the researcher desires to do so.

Survey. A survey refers to information collected by the administration of a test or tests at the present time; that is, at time zero in the simulation. (Each RFP package provides a definition of time zero.) Each time a survey is conducted, the test(s) are re-administered. Individual research subjects will get somewhat different scores in different surveys because of errors of measurement.

Treatment. In FEHR, the term treatment refers to a subroutine which changes a subject's test scores over time. Each treatment is identified by a two-digit code. Variables change over time at different rates depending on which treatment is administered. Tests can be administered at any time during a treatment from zero up to the maximum time available in a problem. The meaning of time zero, the size of a unit of time (e.g., weeks, months, years) and the maximum time available are all specified in the RFP package.

Treatment Group. A treatment group is a set of simulated subjects who receive exactly the same treatment and tests. Note that it *may* contain several "groups" from a research design. For example, if sex and race were design factors, males and females of all races could be included in the same treatment group. This would be advantageous in cases where it costs more to administer the treatment to four small groups than to a single large one.

Card. A card refers to a single line of instructions for the computer such as might be printed on a single punch card. Such an instruction usually contains letters, words, and/or numbers which must be in particular positions in the line. Each card has only a limited number of positions: the maximum number of characters -- including letters, numbers, *and blanks* (or spaces) -- on any one card is 80. Cards must be entered to a computer in a specific order which is prescribed by the instructions which follow.

Column. A column refers to the position from left to right on a card or line. The first position is column 1, the second column 2, and so on. Columns are exactly equivalent to spaces on a typewriter. Each column may contain any legal typewriter character -- including a blank. Note that a blank is entered with a typewriter space bar.

Line. A line refers to a position on a private computer file. This can be thought of as successive lines on a printed page. A line in a file is an exact counterpart of a card. Again, the lines must appear in the order specified by the instructions.

Step. A step refers to a group of cards or lines which are related to the same computer operation. For example, all the cards listing the ID's of the subjects in a particular sample are in the same step.

Steps in Generating Requested Data

To run a request it is necessary to have each of the following program components available: the data generator's main program, the data generator problem packet for the problem being used, and a random access file designated INT which contains the nine blocks of internal Z scores (450 lines with 53 scores per line). Detailed directions for implementing the program at a new site are provided in the game managers manual. The process by which the program generates the requested data is described below in order of request steps.

Step One. The first step in a request deck consists of a set of one or more cards which set a series of program switches or keys to control the operation of a variety of optional features built into the FEHR data generator. The card or cards for step one are prepared by the game manager and supplied to each team prior to their first computer run. The operation of each key is described by the excerpt from the game manager's manual which appears in figure 2.1.

Step Two. The second step of the request deck consists of one card which identifies the data generation subroutine to be used -- file search, survey, or treatment -- and provides the program with a random number parameter. This parameter defines the starting place in the generation sequence. If identical requests are run with identical random number parameters, they yield identical results. If the parameters are not identical, the results will differ by randomly generated measurement errors.

STEP 1. CARD 1

Column	Key	Operation of Key (Note: A blank can be substituted for 0.)
1	0	No summary statistics are printed.
	1	Means and SD's for all variables are printed at the end of each new file search, survey, or treatment.
	2	Means and SD's for all variables are printed at the end of each sub-unit within a file search or survey. Not used for treatments.
2	0	Only the regular printer output (device 6) is obtained. (These are not suitable for direct analysis because they contain many titles and the like.)
	1	Data is output on computer file device 7 as well as on the regular printer (device 6). This permits the user to punch the device 7 results on cards or store them on disk, and still receive a printed output from device 6. The results from device 7 do not contain the headings or titles, and are in a more compact format than device 6 output.
3	0	Variable headings are printed for each new sub-unit within a file search or survey, but (as above) only once for each new treatment.
	1	Variable headings are printed only at the beginning of each new file search.
4	0	The costs are accumulated and continuously compared with the maximum budget entered by the team. When the charges exceed the budget by 5%, the request is aborted.
	1	Costs are computed as above, but there is no abort if the budget is exceeded.
5	0-9	This is the number of subject ID's to be entered on each line. Normally a key of 0 (zero) is used to indicate ten ID's per line. However, you would use a 1 here if the card outputs from a previous survey or file search were to be used to identify the subjects in a subsequent run. Any other number between 2 and 9 can be used if the local devices require it, but note that blank ID's are ignored.
6	0	Variable scores will be returned for each legal ID entered.
	1	Some students will drop out or move within each survey or experimental treatment (but not in a file search). Different students will drop out for each different computer run, but the proportions will be about the same. These proportions (probabilities of attrition) do vary from problem to problem, however, and from group to group within problems.
7	0	The usual built-in treatment effects are in operation.
	1	The built-in treatment effects are each multiplied by a signed decimal constant. This allows the Game Manager to magnify or decrease differences among treatments. It would normally be used only when there were strong pedagogical reasons for changing the treatment effects. If a 1 is entered in column 7, a treatment multiplier card must follow immediately after the key card.

STEP 1. CARD 2

This treatment multiplier card is necessary only if a 1 appears in column 7 or the key card. It contains a multiplier for treatment 1 in columns 1-4, for treatment 2 in columns 5-8, for treatment 3 in columns 9-12, and so on up to the total number of treatments. Each multiplier should be a decimal number.

Figure 2.1. Function of the Keys (or Switches) Entered in Step One of a Request.

Both the request deck and the generation algorithms are different for treatments than for surveys and file searches. In this description, steps three to eight describe file searches and surveys and steps nine to thirteen describe treatments.

Steps Three through Eight: File Searches and Surveys

Step Three. The third step consists of one card which defines the maximum amount of money which can be spent on this file search or survey. Unless the cost default switch was activated in step one, the computer will accumulate costs as each subject is tested and compare the accumulated total with the maximum budget each time. When the total exceeds 105% of the maximum, the computer prints a message that the budget was exceeded then aborts the run.

Step Four. The fourth step consists of one card containing a list of up to twenty three-digit variable indices. These are checked for validity by the program then stored in memory for future reference.

Step Five. The first card in step five contains a key which determines whether the "secretary" subroutine is to be invoked. If the key is zero, the computer proceeds with the next step. If it is one, the computer reads and stores a series of Boolean statements to be used in selecting the subjects. Any of the operators for a FORTRAN logical IF statement may be used to define the desired subjects in terms of their scores on the variables requested in step four. For example, a Boolean statement could be used to print only those subjects who scored less than 2 on the first variable and more than 3 on the second variable. Comprehensive instructions for the use of the secretary feature are provided in the IST unit *Using the FEHR Secretary*.

Step Six. Step six consists of one card defining the number of *cards* in step seven.

Step Seven. Step seven consists of one or more cards containing the ID's of the subjects for whom data is to be returned. These are checked for validity, then stored in memory for future reference.

Generation Process. The program now has sufficient information to proceed with the generation of variable scores for each ID listed. This is accomplished by the search and conversion subroutines.

The search subroutine uses the three part ID to attach one of the 45^0 lines of the internal data base to each individual. Each unit-sub-unit combination is referenced to one of the nine internal blocks. A particular line within the block is then selected by a procedure that ensures that the same line is always associated with the same individual and that a line is not reused until all other lines in the block have been used. The line of internal scores is read from the random access file and then modified according to the sub-unit parameters stored in memory. Four other parameters (M1, M2, M3, M4) are computed and used to control the conversion from internal scores to external scores. The parameters are pointers used by a conversion subroutine: their function is described in a subsequent discussion of that routine. These four parameters are stored separately for each sub-unit to permit control of intergroup differences.

When the appropriate line of INT has been attached to each ID, the four parameters and the set of internal scores (ZVAR) are written in normal deviate form onto I/O device #3, which is a sequential scratch file. The full contents of this file are subsequently used by the other algorithms. Once this sequential file has been constructed for a particular request, the internal data base is not referenced again and the storage allocated to it can be released.

The conversion subroutine generates the external scores requested by the player from the internal Z scores stored on I/O device #3. Each external variable has associated with it a number of parameters that are used in this conversion process. One parameter indicates whether the variable is nominal, ordinal, or interval, and whether it is available on file searches, surveys, and/or treatments. Should it be unavailable for the routine requested, then a value of -99 is returned as the variable score. Another parameter indicates which is the appropriate internal variable to use for the conversion. A third parameter indicates for each external variable classified as interval the population mean for the particular unit-sub-unit combination referenced, while the fourth parameter indicates the standard deviation of that population. Unreliability in the external scores is introduced by using the fifth parameter, reliability, in the manner previously described. In the case of a survey, a different error component is generated each time. For file searches the random number generator is preset to the same value every time the same variable is requested for a particular individual. Thus, the same score is generated each time a file search is conducted of that individual. File searches do not return errorless or "true" scores -- they simply use the same measurement error each time.

Through all of this there is the possibility that the random number generator will cause the external score to be too large or too small to be practically feasible. It was necessary to check each score to make sure it did not exceed the maximum or minimum values established from the test manuals or from the literature. If an external score exceeded the maximum, it was set equal to the maximum parameter. Similarly, if it fell below the minimum, the score was set equal to the minimum.

In performing an ordinal scale conversion, the standard normal distribution was divided into ten equal probability regions and a certain external score was attached to each region. The internal Z score had unreliability added to it in the same manner as for continuous variables. Then the modified internal Z score was examined to see in which region of the normal curve it fell and the attached external score was then used as output. For ordinal variables, only the first three parameters were necessary, with the third being a pointer to the appropriate vector of ten external scores.

Many of the variables usually considered nominal were capable of being generated by the ordinal procedure described above. However, certain critical nominal variables (e.g., sex and race) were stored directly in the internal file. These stored values were used as the external scores of the variables without conversion. Consequently, only the first parameter was necessary for variables of the nominal type.

A technique to obtain a substantial amount of data compression throughout the conversion process was to associate with the unit-sub-unit portion of each ID a set of pointers to the previously discussed parameters rather than the parameters themselves. For example, in many problems there is a need for vastly different variable means from subgroup to subgroup. Rather than have the mean for each variable attached directly to the sub-unit thus many sub-units could be pointed to the same vector of means just by repeating a single digit. This procedure was also followed for the standard deviation, reliability, maxima, and minima parameters. The pointers were attached to each individual depending on his sub-unit membership in the search subroutine, and became part of the sequential scratch file on I/O device #3.

Step Eight. The last step in a file search or survey consists of a single card which contains 'END' if the run is now complete, 'MORE' if more than the maximum number of variables (19) was desired, and 'NEW' if a second request (e.g., one or more treatment groups) are to be obtained on the same run.

Steps Nine through Thirteen: Treatments

Step Nine. Step nine is actually the third step in a treatment request. It consists of one card which identifies: (1) the number of cards of subject ID's, (2) the number of measurement or observation times, (3) the number of different treatments to be administered to these subjects.

Step Ten. Step ten consists of one or more cards listing the ID's of the subjects in this treatment group (i.e., all getting the same treatment).

Step Eleven. Consists of one card for each measurement time. Each measurement card begins with a two digit number identifying the time at which the measurements were taken, followed by a list of three digit numbers identifying the variables to be measured at that time.

Step Twelve. Following the measurement cards are one or more treatment cards. Each of these list two things: the time at which the treatment is to begin and the index of the treatment to be administered.

> Generation Process. The concept of external treatment was implemented by modifying the internal variable rather than operating on the external variables directly. It was necessary to follow this procedure so that external scores that ought to covary would continue to covary after the treatment was applied. For instance, the treatment ought to affect the general ability to read rather than just one particular reading score.
>
> The modification of the internal variables due to a particular treatment is accomplished by the treatment subroutine which consists of two parts. First, each individual's treatment time is modified. Second, the amount of growth on each variable is calculated from the appropriate stored parameters and the modified time parameter. This has the effect of creating a different growth curve for each simulated subject. In calculating the time

modification, the time modification is determined by the initial values of the internal Z scores of the individuals. Then the weighted sum of internal Z scores is computed and divided by the sum of the weights. Since the mean value of the internal Z scores is by definition zero, the mean value of the weighted sum divided by the sum of the weights is also zero. This ratio will be larger than zero whenever the average value of the heavily positively weighted variables was larger than zero and the heavily negatively weighted variables was smaller than zero. The larger the scores of the positively weighted Z scores, the larger the ratio will be and similarly for the negatively weighted Z scores. The time specified by the player is multiplied by this ratio, and the result added to original time to produce a modified time. The modified time will be larger than the player specified time for those individuals that are superior in the positively weighted variables and/or inferior on the negatively weighted variables. It will be smaller than the player specified time for those individuals that are inferior on the positively weighted variables and/or superior on the negatively weighted variables.

After the modified time is determined, a treatment effect is computed. This effect is computed as the amount of change in the initial internal Z score. It is achieved using a set of treatment curves that are specified in the following manner. For each of the internal variables that can be changed there is a unique curve specified for each possible treatment. The curve is the amount of Z score gain as a function of modified time. The curve in all instances is stored as a set of four parameters which can specify a large number of different treatment curves of the type required. Technically, the parameters specify a set of possible transformations to the hyperbolic tangent curve. The hyperbolic tangent was chosen purely because of its convenience in allowing a large number of reasonable treatment curves to be specified.

When all of the internal variables have been modified using the modified time parameter, the full set of internal Z scores is then written out on another sequential scratch file for use in the conversion subroutines from internal to external scores. That is, once an internal score is changed, it becomes the operative internal score for the individual. We choose to operate on a copy of the internal data base rather than on the internal data base directly because one line of the latter can represent many individuals and since different individuals can be put into different treatments, then it is essential that the copy be used rather than the original.

Step Thirteen. Step thirteen consists of a single card which specifies 'END' if the run is completed, 'MORE' if another treatment group from the players' experiment is to be generated, and 'NEW' if another run is to follow (e.g., two different experiments).

SECTION II. DEVELOPING A PROBLEM PACKET

It was mentioned previously that the FEHR computer program required a separate problem packet for each of the eight problems. Each packet specifies a unique set of program parameters which define the research population, the educational treatments, and the variables to be used in the problem concerned. The purpose of this section is to describe the process by which these problem packets were developed. Our intent here is to provide just enough descriptive detail to give the reader an understanding of the scope and complexity of the problem development task.

There are three main phases in the development of a problem packet: (1) operationalizing the problem, (2) specifying the program parameters, and (3) preparing the problem packet (card deck) which defines these parameters for the computer. Each phase is described under the appropriate heading below.

Phase 1: Operationalizing the Problem

The task of preparing a specific operational statement of the problem to be simulated is similar to the specification of a problem to be researched. The major difference lies in the amount of delimitation required. Our procedure consisted of the following ten steps.

1. The content of the problem was identified, and delimited to a researchable scope and size.
2. The literature was searched for previous research relevant to the problem area. A summary of each relevant study was prepared and put in the Information Bank.
3. The problem to be developed was then defined operationally as a discrepancy between what is happening in the (simulated) system and what should be happening in the system.
4. A finite set of possible treatments for remediation of the problem was identified. A verbal description of each treatment and its anticipated effects was then prepared. In general, it was impractical to have more than ten treatments in a problem. (The current educational practice was always included as one of the treatments.)
5. The nature of the research population is specified, and each of the sub-groups within it. The nature of the differences between sub-groups was then described in detail. A sub-group is here defined as any set of subjects who may be expected to have different initial score patterns or to respond to treatments in a unique way.
6. The set of dependent variables (tests) which were to be available in the problem were selected. This list included any variable which should change as a result of a treatment. Next, for each variable, the population mean (M), standard deviation (S), and reliability (R) was defined. In the case of standardized tests, the published statistics were used for this purpose. Otherwise, arbitrary values which appear reasonable to describe the population were chosen. Where multiple statistics were necessary (e.g., when various grade levels had different means and standard deviations) a table of statistics was prepared.

7. The set of independent variables which were to be available in the problem was selected. This list included any variable (test score) which should not change as a result of a treatment; e.g., sex, race, sibling position, SES, and (frequently) IQ. For all except the nominal variables the population mean (M), standard deviation (S), and reliability (R) were defined. Again, the published norms were used for standardized tests and arbitrary values were chosen for the remaining variables. As above, tables were prepared whenever multiple statistics were required.

8. Both the independent and dependent variables were organized into clusters which seemed to be measuring approximately the same thing. Each variable within a cluster was expected to react to treatments in a similar manner. Thus it was possible to generate all the variables within each cluster from the same internal variable. The goal was to have as few internal variables as possible -- particularly criterion (dependent) variables. In subsequent discussion, we shall call these internal references the construct variables. In general we aim to have between ten and twenty constructs, excluding any of the built-in nominal variables. To save space, a number of unimportant variables were clustered about a "garbage" variable and made to appear different by giving them low reliabilities.

9. The sub-groups from step 5 were now clustered into sets having similar minority/majority (e.g., black/white) distributions. All sub-groups within each cluster can thus be referenced to the same internal block.

10. The problem specifications developed above were reviewed to determine whether further delimitation was necessary. The upper limit for the total parameter set is roughly defined by the expression

$$NSG[9(NEV) + 6(NIV + NDV) + 4(NT)(NDV)]$$

where NSG is the number of sub-groups, NEV is the number of external variables, NIV is the number of independent-variable constructs, NDV is the number of dependent-variable constructs, and NT is the number of educational treatments. As a rule of thumb, the problem was further delimited whenever the value of the expression exceeded 10000.

Phase 2: Specifying Program Parameters

The program parameters for a problem were specified by filling out a series of thirteen forms. Facsimilies of these forms are reproduced on the following pages. Cross-check information regarding the preparer, the preparation date, and date of entry into the computer was included on each form, but has been omitted from all but form 1 here to save space. The information required by the forms was recorded in the following steps.

1. The clustered variables from step 8 and the corresponding means, standard deviations, and reliabilities were entered in the blanks to the left of the double line on form 1. The members of each cluster were entered contiguously, and horizontal lines were drawn to separate the clusters of variables from one another.

2. One of the construct variables from internal data base in the file INT was chosen to represent each cluster of external variables. A complete listing of the 53 construct variables in INT appears in figure 2.2. The placeholder names of the variables were used as a guide in this process, but these were not considered definitive labels. Also, the correlation matrix for INT was used as a guide in selecting construct variables. But it was only necessary for the correlations to roughly approximate their theoretical values (i.e., within .25), since it was possible to adjust correlations by means of the Z-add device. This process will be discussed in a subsequent section.

3. A list similar to that in figure 2.2 was used to check off which of the INT variables were to be used in the problem. These were then entered in ascending order on form 2 in the first and third columns. Since only the constructs actually used are read from INT in any one problem, the

COMPLETED BY: ____________________ ON __________ PROBLEM: ______________

DATE ENTERED BY: ________________ ON __________ PAGE ____ OF ____ DRAFT _____

FORM 1: DEFINITION OF EXTERNAL VARIABLES

VARIABLE NAME	VARIABLE NORMS			EXTERNAL INDEX	4 LETTER CODE NAMES			CONSTRUCT INDEX	VARIABLE TYPE
	M	S	R		LINE 1	LINE 2	LINE 3		

FORM 2: TREATMENT WEIGHTS FORM

NEW CONSTRUCT INDEX	CONSTRUCT DESCRIPTION	VARIABLE INDEX IN BLOCK	TREATMENTS						
			1	2	3	4	5	6	7
1									
2									

FORM 3: VECTORS OF MEANS

EXTERNAL INDEX	NAME	'AMEAN'												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1														
2														

FORM 4: VECTORS OF STANDARD DEVIATIONS

EXTERNAL INDEX	NAME	'SD'												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1														
2														

FORM 5: VECTORS OF RELIABILITIES

EXTERNAL INDEX	NAME	'RELIABILITIES'												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1														
2														

FORM 6: VECTORS OF MAXIMUM VALUES

EXTERNAL INDEX	NAME	'MVMAX'												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1														
2														

FORM 7: VECTORS OF MINIMUM VALUES

EXTERNAL INDEX	NAME	'MVMIN'												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1														
2														

FORM 8: DEFINITION OF SUBPOPULATIONS

				SEARCH SUBROUTINE PARAMETERS						
GROUP	(UNIT	SUBUNIT)	N	BLOCK	MEAN	S.D.	MAXMIN	REL	Z-ADD	PRIME

FORM 9: DEFINITION OF SAMPLE SIZES FOR UNIT/SUBUNIT COMBINATIONS

SUBUNIT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1																			
2																			

FORM 10: VECTOR DEFINITION FOR QUASIDISCRETE VARIABLES (See Form 11 for vectors per se)

VARIABLE Z	INDEX	TITLE	CATEGORIES	INDICES OF VECTORS USED

FORM 11: VECTORS USED FOR QUASIDISCRETE VARIABLES

VECTOR INDEX	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1										
2										

FORM 12: Z-ADDS FOR CONSTRUCT VARIABLES

NEW CONSTRUCT INDEX	NAME	Z-ADD INTERNAL											
		1	2	3	4	5	6	7	8	9	10	11	12
1													
2													

FORM 13: DEFINITION OF TREATMENT EFFECTS FOR TREATMENT______

NEW CONSTRUCT INDEX	DESCRIPTION	X	Y	ORIGIN	ROTATION	CURVE
1						
2						

VARIABLE INDEX IN BLOCKS	PLACEHOLDER VARIABLE NAME	VARIABLE INDEX IN BLOCKS	PLACEHOLDER VARIABLE NAME
1	Perceptual Motor	34	Reading Achievement
2	Visual Comprehension	35	Oral Reading Skill
3	Physical Science Achievement	36	Language General
4	Social Studies Achievement	37	Garbage Variable
5	Arts & Humanities Achievement	38	SES Level
6	Health Science Achievement	39	Writing Ability
7	Attention Span	40	Problem Solving Ability
8	Peer Interaction	41	Reasoning Ability
9	Class Participation	42	Perceptual Ability
10	Willingness to Work	43	Age
11	Visual Memory	44	Visual Perception Skills
12	Spatial Ability	45	Auditory Perception Skills
13	Visual Motor	46	Personality Type
14	Auditory Memory	47	Motor Ability
15	Auditory Comprehension	48	Memory
16	Balance & Posture	49	Health
17	Word-Study Skills	50	Sex: 1 = Female
18	Attitude to Major Area		2 = Male
19	Human or Social Skills	51	Ethnic Group: 1 = Majority (white)
20	Mathematical Concepts		2 = Minority (black)
21	Arithmetic Computation	52	FMLY RLTN: 1 = only child
22	Non-verbal Aptitude		2 = oldest of 2
23	Word Approach Skills		3 = youngest of 2
24	Word Comprehension		4 = oldest of 3-5
25	Readiness to Learn		5 = middle of 3-5
26	Verbal Aptitude		6 = youngest of 3-5
27	Total IQ		7 = oldest of 6+
28	Parents Occupation		8 = middle of 6+
29	Parents Education		9 = youngest of 6+
30	Present Achievement (General)	53	Family Marital: 1 = both parents
31	Classroom Behavior		2 = divorced parents
32	Attendance		3 = mom dead
33	G.P.A. (Past Achievement)		4 = dad dead
			5 = step-parent

Figure 2.2. Summary of the Internal or Construct Variables Contained in the INT File.

number in the second column now becomes the construct index for the problem concerned. These indices were entered under the "construct index" heading on form 1.

4. At this point, the material on form 1 was frequently reorganized so as to group variables by their external meaning. For example, all aptitude tests could be grouped together, regardless of their internal construct. When the reorganized material had been copied to a new form, the remaining columns of form 1 were completed as follows.

    (a) The external variable indices were developed by numbering from the top down in the first column to the right of the double line. (Note: All material to the left of the double line is real-life information, while that to the right defines the state of affairs inside the computer.)

    (b) The four-letter codes columns were filled with mnemonic aids to variable recognition. These three codes will be printed by the computer on three lines, one below the other.

    (c) The construct indices were checked for accuracy.

    (d) The "variable type" column was completed according to the key given below. (Note: A variable was defined as quasi-discrete whenever the normal deviate score was to be transformed to an ordinal or nominal scale. For example, parents' education or SES might be generated as quasi-discrete variables in order to control the percentage of cases in each category. Discrete variables are those stored directly; that is, variables generated from variables 50 through 53 in figure 2.2.

KEY

1 = Continuous
2 = Quasi-Discrete — Variables available on file search only.
3 = Discrete

4 = Continuous
5 = Quasi-Discrete — Variables available on file search, survey or treatment.
6 = Discrete

7 = Continuous
8 = Quasi-Discrete — Variables available on survey and treatment.
9 = Discrete
10 = Continuous
11 = Quasi-Discrete — Variables available on treatment only.
12 = Discrete

(e) The variable and external indices from form 1 were copied into the first two columns of forms 3, 4, 5, 6, and 7.

5. The sub-groups of the research population were organized into units and sub-units according to two external classifications. These were usually physical groupings such as schools (units) and classrooms within schools (sub-units). These were listed on form 8 in any logical order which maintained the contiguity of sub-units all belonging to a particular unit. The "group" column was used to verbally define the external name of the group. The "unit" column was filled in by entering a "1" for all groups in the first unit, a "2" for all groups in the second, and so on. The sub-unit column was then completed by numbering continuously (form 1) within each unit.

6. An arbitrary sample size was then defined for each group, and entered under the "N" heading on form 8. For convenient computer entry, this information was then duplicated on form 9.

7. Each sub-unit was assigned to a block according to the desired majority/minority proportions. The appropriate index was then entered in the "block" column of form 8 according to the following key.

BLOCK	% BLACK POPULATION
1	1%
2	10%
3	25%
4	40%
5	50%
6	60%
7	75%
8	90%
9	99%

8. Systematic differences in patterns of scores from sub-group to sub-group were created by defining a Z-add for each construct variable within each group. A Z-add is a constant which is added to the nominal deviate "true" score before an external score is generated. Thus a Z-add of + .5 would raise each subject's true score and the group mean score by half a standard deviation. A different vector of Z-adds was defined for each cluster of sub-groups developed in step 9 of the operationalizing phase. A Z-add values for the first cluster was entered in column 1 of form 12 for each construct variable. Then values for the second cluster were entered in column 2, and so on. When this was completed, the appropriate vector indices were entered on the Z-add column of form 8. Groups referenced to different vectors exhibited different patterns of true scores.

   A second function of the Z-add variables which is less obvious is that they also influence the intercorrelation among constructs. Although the procedure is in practice very complex, the idea is simple. Whenever it was desired to increase correlations, identical amounts were added to all the target variables within each vector, and the differences between vectors were emphasized. When reduced correlations were desired, the Z-adds of the target variables were given opposite signs.

9. Vectors of means for form 3 were developed as follows. First, an "expected mean score" for the first sub-group listed on form 8 was entered in column 1 for each external variable of the continuous type, and a -1 for each variable which is discrete. For each quasi-discrete variable an index number was entered: this index referred to one of the quasi-discrete on form 11.

   Form 11 was developed by the following process. When the first quasi-discrete variable was encountered, the first vector (line) of form 11 was completed by entering in each column the value to be printed if the internal Z score was in the decile indicated by the column heading. A "1" was then entered in column 1 of form 3 to indicate that the first quasi-discrete vector was used. When the second quasi-discrete variable was encountered, the "1" index could be re-used if it defined the desired output, or a new quasi-discrete vector could be defined and the index "2" entered on form 3. This process was repeated until all the quasi-discrete variables were indexed.

At the end of the above process, each variable listed on form 3 had either a mean value or an index number in column 1. At this point, a "1" was entered in the MEAN column of form 8 opposite the first group to reference it to the first vector of means.

If the next group was expected to have a similar pattern of external scores, a "1" was also entered opposite group 2. Otherwise, a second column of means was developed and a "2" was entered in the MEAN column to indicate that this group used the second vector of means. The first step in defining the means of each new group was to examine the existing vectors for a suitable pattern. If one was found, its index was used on form 8. Otherwise a new vector was defined on form 3 and its index was recorded. This process continued until a MEAN index had been defined for each group on form 8.

10. A similar process was used to define the fectors for standard deviations, reliabilities, maximas, and minimas. However, in each of these vectors all quasi-discrete and discrete variables were given a value of zero. At the end of this step, forms 3 through 7 and all but the last column of form 8 were complete.

11. The PRIME column of form 8 was completed by entering a different prime number for each group. These numbers were used by the data generator to define a starting point in the random number generator; this was used to ensure that a file search retrieved the same variable scores for a particular individual each time he was referenced.

12. Form 10 was completed at this point to provide a cross-reference and help eliminate referencing errors. None of the information on this form was new, it was all contained from the other forms. All errors exposed were corrected immediately.

13. The general treatment effects were defined on form 13. First, a separate form was prepared for each treatment by copying the construct indices and descriptions from form 2. Next four parameters were selected for each internal variable so as to define a general growth curve for the variable which was judged to be consistent both with previous research and with the pedagogical purposes of the problem. Several hundred different growth curves had been generated by trial and error using a wide variety of parameter combinations. Once a desirable curve had been selected, we simply entered the parameters associated with the trial curve most like the one desired. At the end of this process, a different growth curve had been defined for each construct variable for each of the treatments.

14. The interactive or moderating effects of variables upon a treatment's learning curves were defined by completing the treatment weight vectors on the right hand side of form 2. The number in the column headings refer to treatment indices: the first column of weights will be used with treatment 1, the second with treatment 2, and so on. These weights were selected according to the following guidelines.

    a. An arbitrary weight between -99 and +99 was entered for each internal variable used for the treatment concerned.

    b. A positive weight implies that the higher the student's initial score on this particular internal variable, the further to the right he will move on each of the learning curves. A negative weight implies that the lower a student's score on this internal variable, the further to the right he will move on each of the learning curves. In both cases, scores at the opposite extreme produce leftward movement.

c. The relative size of each variable weight determines the importance of that variable in moderating the treatment effects.

Phase 3: Preparing the Problem Packet

The card deck defining the problem packet for the problem was prepared by punching a Fortran data statement for each of the vectors defined by the thirteen forms. Although this was a time consuming task, it was largely mechanical, and need not be elaborated here. Once these packages had been successfully compiled, the simulated problem was ready for formative evaluation, which is dealt with in the next chapter.

CHAPTER 3

FORMATIVE EVALUATION AND PROBLEM VALIDATION

This chapter is concerned with the formative evaluation of the entire FEHR system and the presentation of evidence of the validity of the eight problems. The material is organized into three sections. The first section describes the formative evaluation process. The second section chronicles the evolutionary changes to FEHR-PRACTICUM during the formative evaluation. In the third section a series of reports from teams who conducted a variety of simulated research projects within FEHR-PRACTICUM are summarized to illustrate the internal validity and verisimilitude of the eight FEHR problems.

SECTION I. FORMATIVE EVALUATION PROCESS

The formative evaluation of a problem began immediately after the first successful compilation of the problem packet. The first step was to run two successive surveys of a small group of simulated subjects (e.g., a single class) and compute the means, standard deviations and time-time correlations. These statistics were then checked against the input parameters, and adjustments made if large discrepancies occurred.

The second step was to run two extreme groups of subjects (e.g., a high achieving class and a low achieving class) through each of the available treatments taking pre and post measures each time. Since the computer set the classes back to the same starting point at the beginning of each treatment period, this procedure allowed a direct check of treatment effects and their interaction with whatever variable was used in selecting the extreme groups. Again, adjustments to parameters were made if the score-patterns were unsatisfactory.

The third step was to try the problem with a small set of researcher-trainees. Wherever possible we hand-picked these groups to obtain both a high level of skill and a high tolerance for delays, ambiguities and inconsistencies.

The fourth step was a full-fledged trial of the problem in a regular class setting. Although this was technically a summative evaluation of the problem, it was also formative in the sense that comments and criticisms collected during this phase resulted in some of the most important system changes. Since the problems were developed and evaluated sequentially, most of these changes occurred during the evaluation of the first two problems. A summary of the comments of the game managers for the first two off-campus trials are provided to introduce the need for system modifications. We are particularly grateful for the criticisms of Dr. Candy Garrett of Indiana University, Dr. Uldis Smidchens of Western Michigan University, and Dr. William Loadman of Ohio State University. Among the more important comments and criticisms were the following.

(1) Information Bank material was not recent or complete or accurate enough. Specifically:

  (a) Several recent studies were not included in the Bank.

  (b) Validity and reliability information was missing for some variables.

  (c) Several variables have been changed so that FEHR outputs scaled scores, but only raw score statistics appear in the Bank.

(2) Definitions of 1-5 scales (e.g., distractability) need to be made more explicit.

(3) The explanation of the use of the FEHR secretary is totally confusing. Instructions need to be rewritten in a programmed format with illustrations.

(4) Is it possible for the data cards output from a file search or survey to be used to define subjects in a treatment request?

(5) The program was very difficult to debug. Would it be possible to break it down into smaller sections?

(6) The message generator concept was not useful. Students experienced a tremendous information overload without having to deal with teacher strikes.

(7) It would seem more useful to incorporate some of the IST material (e.g., how to prepare a request for the data generator) into the general instructions for the game.

SECTION II. EVOLUTIONARY SYSTEM CHANGES DURING EVALUATION

During the evaluation process needed revisions were made as the need arose. These covered virtually every component of the FEHR system. The substantive changes to each component, and the reasons for each are summarized below.

Introductory Materials

The introductory materials, as conceived in the original proposal, consisted of two separately-bound booklets titled Players' Introductory Booklet and Players' Orientation Booklet. In addition, a"game manager's orientation script" was mentioned. It was to become a section in a comprehensive Game Manager's Manual. The evaluation results and the ensuing revisions are itemized according to those headings.

(1) Players' Introductory Booklet. The prototype booklet produced in 1970-71 consisted of a narrative introduction to the game and an illustrated description of Fair City, U. S. A., the hypothetical city which forms the environment for FEHR-PRACTICUM research. But, in our first full-fledged games during summer 1971, we found that players frequently wished to check information contained in the Fair City description. Consequently, these two sections are now bound as separate booklets. The first is titled Players' Introduction to the FEHR-PRACTICUM Game, and the second Fair City U. S. A.

(2) Players' Orientation Booklet. The prototype version of this booklet was a semi-programmed text in which the players "watched" as a mythical player named Smith completed a sample problem. It was obvious from the first use of these materials during the fall semester, 1972 that drastic revisions were necessary. The orientation

was far too long -- it took almost four hours to complete -- and it required far too much new material for students to assimilate. In particular, two major problems were identified: (1) the emphasis on the terms "game" and "decision" led subjects to search for the "best" research design which they assumed to be hidden in the machine rather than attacking the problem as they would in real life, and (2) players had difficulty learning to use the forms by which they made requests for information.

We experimented with various forms and tried out two nondidactic approaches in our presentations to the Psychology 292 class and the author's Education C650 class. Although the modifications attempted were better received, there were still complaints about the length of time taken for orientation.

We next experimented with a radically different approach. The essence of the new plan was to provide each player with semi-programmed instructions for using each component of the game. The purpose of the orientation in this context was to teach players how to use the instructions. The major advantage of the new format was that players could immediately begin to solve their problem rather than beginning with a sample problem.

The programmed instructions as described in chapter 1 proved much superior. Approximately 200 students have now used this version with little or no problem.

(3) Game Manager's Script for Orientation. The script which accompanied the original orientation was, of course, included in the negative evaluation above, and consequently dropped. Instead, the Manual now contains a series of notes alerting Game Managers to areas which cause students difficulty and suggesting helpful teaching strategies.

The Computer Program

One of the first changes to the computer program was to change the method by which treatments affected test scores. In the initial model, each external variable was changed separately. In January 1972 the program was altered so that treatments modified only internal variables. In the case of PEP, this accomplished a 10:1 reduction in the number of treatment parameters needed and a similar reduction in the number of treatment computations.

During that same period, we had encountered difficulties in translating literal variables from the IBM format at Michigan to that used by the CDC 6600. This trouble was resolved in April 1972 when the literal statements were all rewritten as Hallerith variables, which are available in similar form on all FORTRAN compilers.

As a result of the comments received during our initial evaluation trials, the task of rewriting the entire program in modular form was begun. This was accomplished by late 1972, but the new form of the program was not used externally until 1973. Implementation at Michigan State University (another CDC installation) was accomplished in a single one-day visit using the modularized version.

Two additional options were added to the program in early 1973. First, in line with a suggestion from Ms. Garret, the program was modified to permit subject ID's to be entered into a request one per line. This permitted the output from previous runs (or before-class runs by an instructor) to be used as input, thus simplifying the sample-definition task. Second, a student mobility factor was added. Each student population was given a "probability of moving" parameter. When this feature is used (at the option of the instructor or game manager) students disappear from the class (i.e., move) during the course of an experiment. This permits students to experience the phenomenon of attrition which plagues educational researchers. (Some schools in cities like Detroit frequently have attrition rates of 75% per year or greater.)

During the first six months of 1974, the entire program was again rewritten to make it more comprehensible to others. Although none of the algorithms were changed, the logical flow was improved, and extensive comments inserted. It is this version of the program which accompanies this report.

I.S.T. Units

It has been our initial plan to use the In-Service Training Units to teach people to play FEHR. In addition, they were meant as a "place-holder" to demonstrate how individualized training materials could later be inserted to teach the substantive content of research/evaluation. However, much of the material originally intended for the IST units was no longer needed after the programmed playing instructions were developed. Consequently, this version of FEHR contains only four units instead of the ten originally planned. However, this does not represent a real restriction of the product, since almost all of the material for the planned IST units is incorporated in the instructions.

The Message Generator

In early descriptions of FEHR-PRACTICUM, considerable importance was attached to the message generator. However, several of the early users of the system thought that players already had too much to contend with and that messages such as notifications of a teachers strike were of marginal pedagogical use. In the current version, the message generator concept is an option which may be used by the game manager, but is not a necessary component of the Practicum. (A variety of suggested messages appear in the Game Manager's Manual.)

The Information Bank

During our first field trials, the Information Bank was roundly criticized for being out of date and for having incomplete summaries of the studies represented. Since it was patently impossible to keep the Bank completely up to date, and since summaries by their very nature are less than comprehensive, we decided that an error had been made in defining the role of the Bank. Rather than defining it as a simulated library, we now describe it as a set of abstracts which may be used either as a placeholder for the literature search (not a substitute), or as a preliminary screening device to determine which materials should be read in their entirety.

RFP Packages (Problem Descriptions)

The early problem descriptions tended to be lengthy narrative descriptions. These were heavily criticized by many of the early users. An additional writing burden was imposed by the decision to develop structured and restricted versions of each problem in addition to the full-fledged original version. Consequently, there was an accute need to provide more compact and succinct descriptions. This was accomplished in two ways. First, a hierarchical organization was developed. Summary information was given first, with supporting details in a variety of appendices. Second, the "checklist of assigned tasks" was developed to permit instructors to define their own problem structures and restrictions. The latter obviated the need for multiple versions of each problem. (The functions of both these devices were fully described in chapter 1, and need not be repeated here.) The present RFP (Request for Proposals) format was adopted for the third problem (Extended School Year), and was so well received that the first two problems were immediately rewritten in that form.

Reduced Number of Problems

The only departure from the original contract specifications which was not explicitly ratified by USOE was the decision to discontinue the development of problems nine and ten in favor of multiple versions of each problem. However, this decision was fully justified, we feel, for several reasons.

The problems, according to the specifications of the progress report dated October 20, 1972, were to feature performance contracts based on the data base in remedial arithmetic and the reading assessment problems (2 and 5). However, after three months of exploratory work, we had failed to discover a vehicle for negotiating the performance contract (e.g., a blank contract) which was realistic and comprehensive without being prohibitively didactic. In addition, the first programming of the READ problem had proven woefully inadequate: the entire development process had to be repeated. All of these events favored the discontinuance of problems nine and ten.

But the final decision was based on cost effectiveness arguments. It was, we decided, a better use of time and resources to develop three different versions (restricted, structured, and unstructured) than to pursue a problem content which offered little chance of success. The triple version alternative had the additional advantage of broadening the appeal of the total package, since less sophisticated students could benefit from the simpler problem versions. This turned out to be a particularly fortunate decision: the "checklist" format eventually developed resulted in not three but many levels of difficulty/complexity for each problem. (The exact number is, of course, finite but undeterminate.)

SECTION III. ILLUSTRATIVE FEHR PROJECT REPORTS

In this section we shall attempt to illustrate the validity and verisimilitude of each problem. It is assumed that the best evidence of validity and verisimilitude is the product which resulted. Consequently, we shall present one illustrative project report for each problem. Following each illustrative report is a brief summary of the evaluative comments made by trainees and game managers who have used the problem. Since these are intended to reflect the present status of the problems, comments which are no longer apropos (e.g., because of changes to the system) have been omitted. Since space prohibits the reproduction of complete reports, only one of these illustrative projects is presented in its entirety.

In the following pages, a sample final report and evaluative comments are presented for each of the eight problem content areas. These are presented in order of their RFP number as outlined below:

RFP001	Perceptual Education Problem (PEP)
RFP002	Remedial Arithmetic (REMAR)
RFP003	Extended School Year (EXTSY)
RFP004	Early Childhood Education (HEADSTART)
RFP005	Reading Assessment Problem (READ)
RFP006	Validation of a Teacher Questionnaire (TQUEST)
RFP007	Remedial Math for Adults (RMA)
RFP008	Busing to Achieve Integration (BUS)

For reasons described in context, only the report for the REMAR problem is reproduced in its entirety. The remainder are summaries of reports at varying levels of sophistication. It is important for the reader to recognize that these are the products of trainees -- many of whom have had no previous research training or experience. Inclusion in this section is not meant to imply unreserved support for either the problem definition or the research strategy.

I. RFP001: PERCEPTUAL EDUCATION PROGRAM (PEP)

The information contained in this summary is based on a report by FEHR trainees 12, 24, and 54. The general assignment was to experimentally evaluate the effectiveness, relative to present practice (PP), of two programs for students in Fair City's elementary schools who are not making satisfactory academic progress even though they are of at least average intelligence and have no serious uncorrected physical disabilities. One of the programs to be evaluated was designated VPM because it featured visual perceptual motor training. The second program was designated SLD to indicate its emphasis on the diagnosis and treatment of specific learning disabilities.

The members of the evaluation team had all completed two semesters of research design and data analysis prior to this project. The material below is a shortened and simplified summary of their final project report.

A. Illustrative Report

Problem. After an extensive review of the literature, the target population was those students in grades 1-3 at Jackandjill school who were scoring in the bottom 15% of the national norm group on the SAT tests for word meaning, paragraph meaning, arithmetic computation, and arithmetic concepts. The problem was to determine which of the three programs -- present practice, VPM, and SLD -- produce the most growth on the four SAT tests listed above.

Hypotheses. On the basis of a careful review of learning theory it was hypothesized that the treatment means would rank PP < VPM < SLD in ascending order of effectiveness at all three grade levels.

Method. A stratified random sample of 120 subjects was chosen from the target population, with 40 subjects in each grade. Educational effectiveness was defined as the average of the four SAT tests when each test score has been transformed to a Z score in the appropriate normative population. The design was a 3 x 3 analysis of variance (grades x treatments), with the average Z score as the dependent variable.

Results. The overall analysis of variance yielded a significant difference among treatments, but there was no significant main effect for grades, and no significant grade by treatment interaction. In subsequent planned comparisons, both VPM and SLD means were significantly greater than PP. The SLD mean was larger than VPM, as hypothesized, but the difference was not significant.

Conclusions. It was concluded that both VPM and SLD programs produced more learning as measured by the SAT tests than did PP. Since the VPM treatment was more expensive than SLD and the latter produced the higher mean score, the researchers recommended that the SLD program be implemented.

B. Summary of Evaluative Comments: RFP001

1. Many users consider this problem to be the most absorbing of the set. Virtually every trainee who has used it mentioned that it strongly motivated outside reading about the tests available (particularly with reference to their reliability and validity), and the original materials from which the information bank entrees were obtained. More than 75% of the trainees reported doing a comprehensive search of the literature even when it was not assigned.

2. The absorption mentioned above sometimes caused difficulties. Several trainees reported annoyance because their favorite tests and/or treatments were not included.

3. Problem definition was found to be particularly difficult in this problem, and disagreements over the nature -- even the existence -- of "perceptual handicaps" were frequently quite vociferous. However, this was considered to be an advantage by some content-area instructors. There was much less contention where the game manager and all teams met to obtain a consensus definition prior to the preparation of final-draft proposals.

4. More than in any other problem, trainees here tended to critize "inadequate" treatments and "invalid" tests. Game managers suggest it is necessary to emphasize that the process is being studied rather than the specific program elements.

5. Because of the effects of comments 3 and 4, above, the Information Bank articles were generally considered an inadequate basis for a project. They were considered a useful stimulus to further reading. It is doubtful whether a meaningful PEP project can be done without a complete library search.

6. The results summarized above are somewhat unusual in that there were significant effects on the achievement variables. Although the model specifies mild positive treatment effects on achievement, these do not generally reach significance with small samples. Larger differences are usually evident on the perceptual variables.

II. RFP002: REMEDIAL ARITHMETIC (REMAR)

The information contained herein is the complete text of a report by FEHR trainee 201. Since summaries are not necessarily representative of the product, it seemed wise to present at least one report in its entirety. The REMAR problem was a natural choice, since it is the standard problem.

The author of this report was a student in C655, the beginning course in research design and data analysis at the University of Michigan. Each member of the class was assigned two treatments to evaluate and asked to choose a moderator variable which he (she) believed might alter the effect of the treatment. The reports were to be written in succinct outline form, and were due on the last day of class. This particular report was not chosen because of its high quality (it ranked in the bottom quartile out of a set of 63) but because it was succinct and brief. Many of the better project reports ran from 35 to 50 typewritten pages in length. One of the required appendices was the budget from the original project. This has been included for illustrative purposes. However, the second required appendix, a log of activities, was omitted because of its excessive length.

A. Illustrative Report

(Title)

REMEDIATION OF 7TH GRADE ARITHMETIC SKILLS VIA AUTOMATH AND IRA

I. INTRODUCTION

Problem

1. This proposal is concerned with the general problem of evaluating two remedial arithmetic programs designed to help grade seven students master the basic computation skills.
2. Large numbers of FEHR City mathematics teachers have complained that a considerable number of seventh grade students are unable to add, subtract, multiply and divide well enough to succeed in the regular curriculum.
3. The mathematics teachers, as a group, have indicated that they believe mastery of simple computational skills is a prerequisite to success in all occupational realms.

4. The societal expectation is that boys will achieve at a higher level than girls on computational skills. Perhaps society views computational skills more crucial for boys for occupational success.

5. Seventh grade arithmetic students in need of remediation will be identified by their scores on the Criterion Referenced Mastery Tests.

6. Purpose of this proposal is to compare the achievement scores of remedial seventh grade arithmetic students who use AUTOMATH or IRA, to those who use the Present Practice.

Definition of Terms

1. Present Practice - students remain in their regular classes.

2. AUTOMATH - students will leave their regular mathematics classes for four one-half hour sessions per week to work with a computer program which automatically administers a series of drill and practice exercises in the basic computational skills.

3. IRA - students will leave their regular mathematics classes for four one-half hour sessions per week to work with a programmed text that administers a series of drills and practice exercises in the basic computational skills.

4. Criterion Referenced Mastery Tests - these tests allow the testing of specific arithmetic computational skills and concepts in addition, subtraction, multiplication and division. Each concept tested presupposes a mastery of the concepts preceding it. A student's success on these tests is equated with complete mastery or 100% correct.

Review of the Research

1. Melson (1971) indicated non-graded instruction of mathematics was an alternative to criterion based evaluations of fifth grade students.
2. Ginsberg (1972) concluded that individualized prescribed instruction may pose immense problems because of children's varied and complex conceptions of mathematics.
3. VanDyke (1972) reported that short intervals of delay in knowledge of results with computer assisted instruction had no significant effect on the learning or test performance of subjects. However, delay of knowledge of results related to poorer attitudes toward computer assisted instruction among women than in men.
4. Maertens (1969) analyzed the effects of arithmetic homework upon the arithmetic achievement of third grade students.
   (a) The results of this study indicated no significant differences in the achievement of groups receiving homework over those not receiving homework.
   (b) Sources of experimental invalidity included;
      (1) Selection (internal) - entire classrooms of students were captive groups.
      (2) Statistical analysis was not shown and chi square was apparently not used. An Fmax reading would have been helpful.
      (3) Within group statistical analysis might have been valuable to develop trends.
5. Summary - The evidence in favor of computer assisted instruction and individually prescribed instruction is unclear. However, it is certain that careful considerations must be given to the type of learning task

and individuals involved. In addition, delay in knowledge of results, occurring perhaps from a breakdown in machinery may result in negative attitudes toward computer assisted instruction.

Conceptual Framework (rationale)

1. Because of the sequencing of steps in learning mathematics it is reasonable to believe that IRA and AUTOMATH will be conducive to remedial progress.
2. The design of AUTOMATH indicates student independence from reading skills. This feature should be to the advantage of students with reading problems.
3. The novelty and motivational aspects of working with a computer should be to the advantage of students who have difficulty in attending to tasks.
4. It has been the author's experience, as a classroom teacher, that males and females in the middle grades achieve in arithmetic at about the same level.
5. It has been the author's experience, as a classroom teacher in the middle grades, that students showing achievement in arithmetic computation also exhibit mastery of arithmetic concepts and vice versa.
6. Overall, it appears AUTOMATH may have the effect of boosting computational skills, personal confidence from success, and attitudes towards mathematics.

Hypothesis

Given the achievement test scores (SATCOMP & ITBSCONC), the computationally remedial students who have been trained in their respective treatments will show the following relationships:

1. AUTOMATH scores will be higher than IRA or Present Practice scores.
2. IRA scores will be higher than Present Practice scores.

3. Male and female students across treatments will score at about the same level.

II. METHOD

Subjects

The target population of this study will be 767 seventh grade mathematics students. The sample population will be randomly drawn from the identified computationally disadvantaged students who score less than 100% on the CRTDIV. 120 seventh grade arithmetic students from the John Watts School will serve as the subjects. A random numbers table will be used to select the remedial students. There is no reason to expect that this particular sample will be biased in any way. Thus, it is reasonable to assume that any definitive results can be generalized to the total population.

Treatments

The treatments to be tested consist of three types: AUTOMATH, IRA and Present Practice (control). These treatments were described under the definition of terms section.

Instrumentation

The instruments for measuring achievement are standardized and highly regarded measures:

1. Stanford Achievement Test: Arithmetic Computation - .87 reliability.
2. Iowa Test of Basic Skills: Math Concepts - .98 reliability.
   (a) Abbreviations - SATCOMP; ITBSCONC.

Design

The subjects will be randomly assigned to three treatment groups of size 40, and then the subjects will be sorted by sex. Since the ratio of males to females is unequal, it will not be possible to obtain a perfectly equal distribution of subjects by sex within treatments.

A diagrammatical explanation of the design is given below. The symbol O stands for a set of observations taken at one time. X indicates the treatment. X_1 = AUTOMATH. X_2 = IRA. X_3 = Present Practice (control). Y represents the moderator variable. Y_1 = females. Y_2 = males. R denotes random selection from the total population.

R	X_1	Y_1	O_1	(Females in AUTOMATH)
R	X_1	Y_2	O_2	(Males in AUTOMATH)
R	X_2	Y_1	O_3	(Females in IRA)
R	X_2	Y_2	O_4	(Males in IRA)
R	X_3	Y_1	O_5	(Females in Present Practice)
R	X_3	Y_2	O_6	(Males in Present Practice)

Procedure

At the beginning of the experiment, the first day of school, the CRT's will be administered to seventh grade arithmetic students. The computationally disadvantaged students will be drawn from the CRTDIV scores which are less than 100% correct. During the remainder of the semester (15 weeks), the three groups of forty students will receive one of the three treatments. Group 1 will receive AUTOMATH. Group 2 will receive IRA. Group 3 will receive Present Practice (control). At the end of fifteen weeks, each group will be administered the SATCOMP and ITBSCONC (post-tests).

Experimental Rationale

The design outlined above met all the criteria for a post test-only experiment, and the procedure outlined assures equivalent experimental histories. We can generalize to future seventh grade arithmetic students in FEHR City only if the results are unequivocal.

Analysis

For the purpose of analysis, an analysis of variance of the six sub-groups followed by an Fmax check, F test and t test will be used.

Step 1. Compute an ANOVA of SATCOMP and ITBSCONC scores to test for differences among means after treatments.

Step 2. Test homogeneity of variance (also checks additivity) using Fmax. This is done by dividing the largest sub-group variance by the smallest subgroup variance.

Step 3. If Fmax is larger than the tabled value for the .05 significance level, check the .01 significance level; otherwise use .05 level.

Step 4. If the F test for the overall ANOVA differences among sub-group means for SATCOMP and ITBSCONC is not significant at the level set up in step 3, analysis is discontinued. If the F is significant, continue to step 5.

Step 5. Find out whether the three hypothesis are supported by the data by performing t tests of:

(1) The difference between the means for all subjects using the AUTOMATH treatment, IRA treatment, and Present Practice treatment.

(2) The difference in the mean for all subjects using the AUTOMATH treatment and IRA treatment.

(3) The difference in the mean for all female subjects and the mean for all male subjects.

(4) If there is a difference between the means of males and females, then examine within treatments.

Data Matrix

The I. D. #, sex, SATCOMP score and ITBSCONC score for each subject will be entered on a punch card, as illustrated in the following diagram. The sexes are intermingled, but the three treatments are not. The first 40 cards will be in the AUTOMATH group. The second 40 cards will be in the IRA group. And the third group of 40 cards will be in the Present Practice (control). The format to be used for punching these cards is: (F8.0, F2.0, F3.0).

GROUP		I.D.	SEX	SATCOMP	ITBSCONC
	1	---	---	---	---
AUTOMATH					
	40	---	---	---	---
	41	---	---	---	---
IRA					
	80	---	---	---	---
	81	---	---	---	---
Present Practice	120	---	---	---	---

III. RESULTS

The results are presented in order of the steps outlined in the analysis section. Within each step, information is organized into four parts, the null hypothesis being tested, the result of the statistical computation involved, the statistical conclusion, and the educational interpretation of that conclusion. The subscripts for the null hypothesis identify the observation from which the data comes.

Step 1 - NULL H_0: $u_1 = u_2 = u_3 = u_4 = u_5 = u_6$

RESULT: Analysis of variance of SATCOMP:
Scores N = 120

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	5	1715.7	343.15	10.524	.000
Within	114	3717.2	32.607		
Total	119	5433.0			

GROUP	N	MEAN	VARIANCE	STD. DEV.
1	20	12.850	19.292	4.392
2	20	13.500	17.737	4.211
3	14	22.143	56.901	7.543
4	26	21.692	41.662	6.454
5	21	14.905	31.790	5.638
6	19	14.579	33.146	5.757

STAT. CONCL: There were significant differences among the six sub-group means for SATCOMP (reject null).

ED. IMPL: Either the treatments or sex or an interaction between the two produced a significant effect on SATCOMP.

NULL H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

RESULT: Analysis of variance of ITBSCONC.
Scores N = 120

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	5	174.54	34.908	.825	.534
Within	114	4821.4	42.293		
Total	119	4996.0			

GROUP	N	MEAN	VARIANCE	STD. DEV.
1	20	16.200	20.221	5.405
2	20	12.950	55.734	7.465
3	14	13.000	47.385	6.883
4	26	13.154	44.615	6.679
5	21	14.905	43.900	6.632
6	19	14.684	33.117	5.754

STAT. CONCL: No significant differences among the six sub-group means for ITBSCONC (accept null).

ED. IMPL: Treatments had no apparent effect on mathematical concepts.

Step 2 - NULL H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$ SATCOMP

RESULT: $F_{max} = \frac{56.901}{17.737} = 3.208$

Critical value for df (6,20) = 3.76;
probability of F_{max} by chance > .05.

STAT. CONCL: A significant difference exists among the sub-group variances for SATCOMP scores.

ED. IMPL: The assumption of homogeneity of variance (treatment additivity) was not violated.

NULL H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$ ITBSCONC

RESULT: $F_{max} = \frac{55.734}{29.221} = 1.907$

Critical value for df (6,20) = 3.76; probability of F_{max} by chance > .05.

STAT. CONCL: No significant differences among the sub-group variances for ITBSCONC scores.

ED. IMPL: The assumption of homogeneity of variance (treatment additivity) was not violated. The lack of statistical significance among sub-group variances may not be attributed to lack of homogeneity of variance.

Step 3 - F_{max} for both SATCOMP abd ITBSCONC within the > .05 level.

Step 4 - Because both initial differences and homogeneity of variance were supported at the .05 level this degree of significance will be used in all following analysis.

Step 5 - Assessment of the three major hypothesis using t tests on SATCOMP scores.

(1) NULL H_0: $\mu_{(3\&4)} = \mu_{(1\&2)}$

RESULT: See contrast 1 on following page.

STAT. CONCL: There is a significant difference among treatment means.

ED. IMPL: AUTOMATH is superior to Present Practice (control) in boosting computational skills.

(2) NULL H_0: $\mu_{(3\&4)} = \mu_{(5\&6)}$

RESULT: See contrast 2 on following page.

STAT. CONCL: There is a significant difference among treatment means.

ED.
IMPL: AUTOMATH is superior to IRA in boosting computational skills.

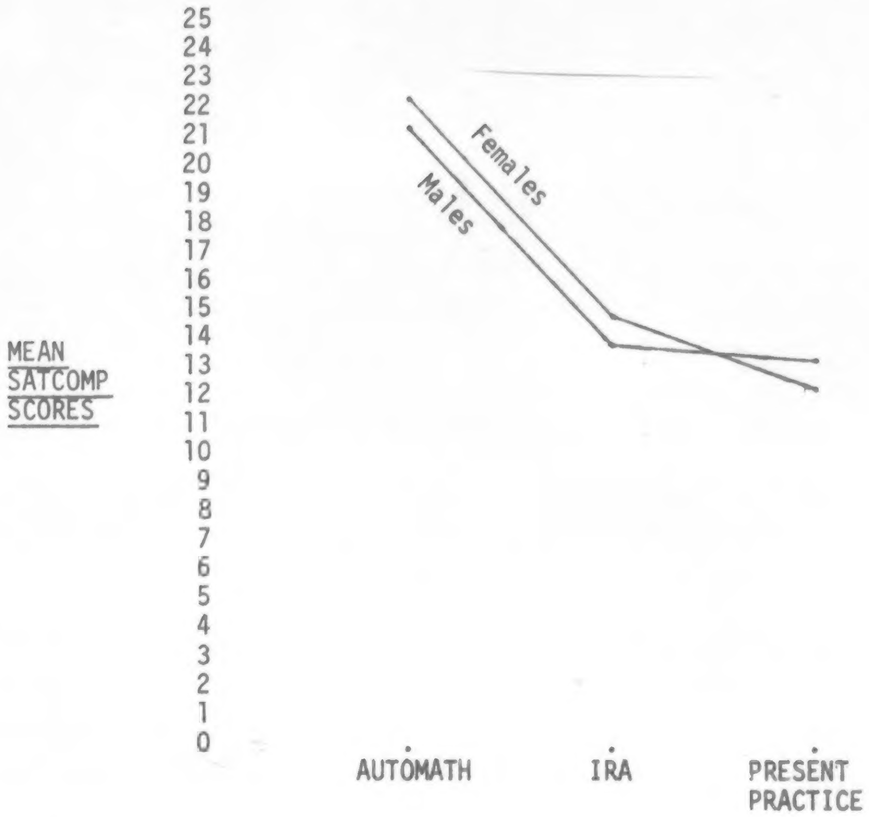
(3) NULL H_0: $\mu_{(1,3\&5)} = \mu_{(2,4\&6)}$

RESULT: See contrast 3 below.

STAT.
CONCL: No significant differences between sex means.

ED.
IMPL: Overall, males and females achieved at about the same level.

	CONTRAST OBSERVED	T-STAT.	SIGNIF.
(1)	17.485	6.683	.000
(2)	14.351	5.482	.000
(3)	-.126	-.039	.968



IV. DISCUSSION

1. Differences among the post-treatment SATCOMP means of six sub-groups were significant. The differences among the post-treatment ITBSCONC means of the six sub-groups were not significant, although homogeneity of variance was indicated.
2. The subjects represented a random sample of the total computationally disadvantaged 7th grade population. This sample is applicable to future 7th grade computationally disadvantaged students.
3. The results of this investigation can be generalized to the total population of present 7th grade computationally disadvantaged students, as well as future 7th grade computationally disadvantaged students.
4. The ITBSCONC scores were not significant indicating a need for future study as to how to boost mathematical concepts along with computational skills.

V. CONCLUSIONS AND RECOMMENDATIONS

1. Among the sample population, the AUTOMATH treatment was more successful than IRA in boosting computational skills.
2. The Present Practice is not a viable method for remediation of computational skills.
3. Males and females performed at about the same level, regardless of treatment.
4. The success of AUTOMATH justifies future cost in adapting it as a course of action to remediate computationally disadvantaged 7th grade students.

REFERENCES

Ginsberg, H. "Children's Knowledge and Individualized Instruction". Educational Technology, 1972, (Mar.), Vol. 12(3), 8-12.

Maertens, N. "An Analysis of the Effects of Arithmetic Homework Upon the Arithmetic Achievement of Third-Grade Students". Arithmetic Teacher, 1869 (My.), 16: 383-389.

Melson, R. V. "Can All Your Fifth Graders Do Page 153? How Haverford Township Solved the Problem: Non-graded Mathematics". Pennsylvania School Journal, N., 1971, 120, 63-67.

Van Dyke, B. F. "Computer Assisted Instruction: Performance and Attitudes". Journal of Educational Research, 1972, 65, 191-192.

B. Summary of Evaluative Comments: RFP002

1. Virtually all users commented on the degree to which REMAR stimulated outside readings of test manuals and critiques, test theory (especially regarding criterion referenced tests), and research methods. Although it did stimulate outside reading of evaluation studies in mathematics education, the trainee enthusiasm was observably less than for the PEP problem.
2. Problem definition for REMAR was within the capacity of most trainees. There was little contention generated; most users seemed to feel the treatments and the treatment effects were sensible and realistic.
3. Game managers and instructors considered the problem especially well suited to training students in rational, objective, and scientifically detached assessment methods. In this respect, there was a sharp distinction between REMAR and such emotionally laden problems as PEP, HEADSTART, and BUSING.
4. The availability of both normative and criterion-referenced tests was considered one of the problem's strengths.

III. RFP003: EXTENDED SCHOOL YEAR (ESY)

The information in this summary is based on a report by FEHR trainees 15, 24, and 62. Since the members of this team had completed two semesters of research design and statistics and had been involved in two previous FEHR projects, their assignment was to assess the overall effectiveness of two experimental approaches to an extended school year relative to the costs of the current program. Although the original report devoted a considerable amount of space to theory development and hypothesis testing, that section of the study which deals with arriving at an objective decision among programs is reproduced in some detail.

A. Illustrative Report

Problem. The aim of this project was to determine the comparative effectiveness and efficiency of three programs: Present Practice, 45-15, and Continuous Progress. The project sought to determine which program increased achievement in language, reading, and mathematical concepts; improved parental attitude toward school program; and did so at the most reasonable cost.

Proposed Remedies. Two extended school year programs were proposed. The 45-15 Cycling Plan retains the regular number of school days (180) but distributes them differently: a repeated cycle of 45 days of school followed by 15 days of vacation. The school is divided into four groups, only three of which are attending school at any one time. The division is on geographical lines so that neighborhoods and families are not disrupted. While the yearly per pupil costs are expected to rise by 8% (9.5% the first year), the net saving over five years should be some 6.5 million, because of the reduction in need for new buildings. In addition, the shorter vacation periods should improve learning by reducing the demands on information retention.

The Continuous Progress Plan increases the school days to approximately 200, with students attending school in 5 to 9 week cycles with two-week vacations between cycles. The school is divided into five geographical groups with four groups in school at any one period. This plan adds an extra 4.5% to the pupil per year cost (6.5% the first year), but saves 7.4 million over the first five years by obviating the need for new buildings. This plan is particularly interesting in core city areas because it affords the disadvantaged child added time and attention. Both extended school year programs may create some resistance because of their effect on family vacation plans.

Sample. The target population of this study was the entire set of elementary schools in Fair City. Because SES appears to affect achievement of students, attitude of parents, and the degree of overcrowding, the population of schools was stratified into three levels. This stratification was performed on the basis of residence descriptions indicated in the Fair City files on the schools. Three schools were randomly selected from each strata, and each school in a residential strata was randomly assigned to one of the three programs. The stratification and randomization procedures were conducted to increase the generalizability of the results since it considered all strata in the community and allowed their investigation.

For analysis and comparison of the effect of the programs on achievement, attendance, and attitude of parents, thirty students were randomly selected from grades one through five in each school. The sample was limited to these grades to enable comparisons over a two year period on all specified variables. Kindergarten children were excluded because the achievement tests do not apply to this grade; grade six students were excluded because they would leave the school before the two year period was completed. Therefore the total sample consisted of nine schools and two hundred seventy pupils. The mean score of the sample of pupils in each school represents the score for each of the nine schools.

Variables and Instrumentation. Tests were selected to provide valid indices of school achievement in the most basic skills required of elementary school children. An examination of content, reliability measures and relative expense resulted in the choice of the Stanford Achievement Test. One set of dependent variables consisted of achievement as measured by three Stanford Achievement Test subtests on word meaning, paragraph meaning and arithmetic concepts. Other dependent variables included: parental attitude toward program, per pupil cost and attendance. The independent variables were programs (PP, 45-15, CP), and residential strata or school composition (predominantly lower class, predominantly blue collar, predominantly middle class). In addition, family SES was used as a covariate to test its effects on achievement variables and remove its effects from the school and residential composition strata factors.

Design. A 3 x 3 x 3 factorial design with repeated measures on the third factor was the data collection guide. Factor 1 was program. It had three levels. Level A was the traditional or Present Practice (PP). Levels B and C were experimental treatments with B the 45-15 plan and C the Continuous Progress (CP) plan. Factor 2 was residential strata or school factor. Three schools were selected for each program and each represented one of three residential levels. The three schools constituting level A were of predominantly lower class black residential composition; students attending level B schools were from predominantly blue collar residential white, black, or racially mixed areas; students in level C schools were predominantly children of middle class home owners and upper class apartment residents. Factor 3 was time at which variables were measured. There were three levels. Level A was the initial start of the school year (time 00), level B was the spring of the first year (time 01), and level C was the spring of the second year (time 02).

At each time interval, the Stanford Achievement Tests on paragraph meaning, word meaning and mathematical concepts were administered to pupils. Also, at each time interval, parental attitude to the program their child participated in was surveyed and cost per pupil ascertained. At times 01 and 02 pupil attendance was determined. During the initial test administration at time 00, information regarding student's race and socioeconomic status (SES) was sought.

Decision Rule. Prior to the commencement of the project, the team arbitrarily assigned relative importance weights to each of the dependent variables. The major emphasis was put on achievement with a total weight of eight. This was evenly distributed to reading (two tests at 2 each) and mathematics (one test at 4). Per pupil cost, which was considered about half as important as achievement, was given an important weight of 4. Parental attitude and pupil attendance were each given a unit importance weight.

Method. A series of specific hypotheses related to the theoretical advantages of each treatment were developed and tested via analysis of covariance. These are omitted from this summary. The procedure for arriving at a decision is described in the results section.

Results. Mean gain scores for each of the treatments were computed for each achievement score. For the other variables, time two means were used directly. The following table represents the results of that procedure.

TABLE 3.1. MEAN SCORES FOR USE IN WEIGHTING PROCEDURE

	Achievement Gain Scores			Other Variable Averages		
Program	Word Meaning	Paragraph Meaning	Math Concepts	Parent Attitude	Pupil Attendance	Per Pupil Cost
PP	5.9	10.6	1.8	2	3.4	2.7
45-15	8.0	15.5	5.8	4.4	4.7	2.7
CP	4.9	9.0	-.3	3.8	3.6	2.0

The above raw means were reduced proportionally to a score ranging from 0 to 1 by dividing by the largest number in the column. These transformed means were then multiplied by their assigned weights and a ranking determined by finding the total score for each program. The results are shown in the following table.

TABLE 3.2. TRANSFORMED WEIGHTED SCORES USED TO ASCERTAIN RANKING

	Achievement Gain Scores			Other Variable Averages			Program Total
Program	Word Meaning (2)	Paragraph Meaning (2)	Math Concepts (4)	Parent Attitude (1)	Pupil Attendance (1)	Per Pupil Cost (4)	
PP	1.48	1.36	1.38	.45	.72	1.20	7.59
45-15	2.00	2.00	4.00	1.00	1.00	1.20	11.20
CP	1.22	1.16	0	.86	.77	4.00	8.01

Conclusions and Recommendation. On the basis of this approximated tranformation, the project team recommends the implementation of the 45-15 cycling plan which has a composite effectiveness score greater than either Present Practice or the Continuous Progress Program. This recommendation was further supported by the fact that on all dependent variables except cost per pupil, the 45-15 cycling plan was superior to the Continuous Progress Experimental Plan as well as the Present Practice plan.

B. Summary of Evaluative Comments: RFP003

1. The Extended School Year (ESY) problem shared most of the strengths listed for REMAR. It stimulated trainees to individual study of test manuals and critiques, general research methodology, and the literature in general. It did not offer the experience with both normative and criterion-referenced tests.
2. Although the definition of a success criterion was as difficult for ESY as for PEP, it seemed much easier for trainees to make rational and detached judgments here.
3. The problem was especially popular among school administrators who saw it as directly related to the type of decisions they made in their jobs.

4. Both instructors and trainees found the per pupil cost factor both interesting and valuable. However, there was some criticism of its representation as a test "score" for each research subject in a study rather than a school or classroom score.
   [Note: All individuals within a unit receive identical "per pupil cost" scores -- there is no other way to generate unit or subunit scores in our model.]
5. Some administrators who used ESY felt that the clear advantage to the 45-15 plan yielded by our model was contrary to some research evidence. [The FEHR staff discounted this comment, however, since it is possible to define success so that the CP plan comes out superior.]

IV. RFP004: HEADSTART (HST)

The information contained in this summary is based on a report by FEHR trainees 1, 13, and 65. Their general assignment was to evaluate the effectiveness of Fair City Headstart project in overcoming deficits in school performance common to culturally deprived children. The three members of this team had completed a first course in research design and statistics prior to beginning the project, and were concurrently enrolled in the second course. Because the original project was concerned with the relative effectiveness of seven different treatment-teacher combinations as measured by thirteen different dependent variables, the following summary deals with only the broad pattern of findings with respect to the three main compensatory programs.

A. Illustrative Report

Problem. The purpose of this study was to determine the immediate effects of the various local Headstart programs (funded under the aegis of the national project of the same name) on measured intelligence, reading readiness and personality, and to assess the effects of these experiences on reading and mathematics achievement in grade one. The effects of the three compensatory curriculums -- Piagetian, language based, and unit based -- were to be evaluated relative to the effects of the present practice (i.e., staying home). Since the national project was especially targeted on minority groups, the differential effectiveness of the programs (if any) by race and sex was of interest.

Hypotheses. In lieu of a dearth of evidence regarding the relative efficacy of the methods, this was viewed as an exploratory study. Comparisons among treatments were planned a priori, but all other contrasts were considered post hoc.

Method. The target population consisted of all Fair City children who were three years old at the initiation of the study, and who fit the following definition of cultural deprivation (CD): (1) Stanford Binet IQ $\leq$ 90 (first quartile); (2) Deutch SES $\leq$ 2 (low SES); and with no physical or perceptual disabilities (i.e., scores greater than 1 on the health, vision and hearing variables). Eight of Fair City's elementary schools were selected randomly, and their entire populations were surveyed on SB IQ, SES, race, sex, health, vision, and hearing variables. All students who fit the CD definition above were identified. These students were then partitioned by race and sex. From each race-sex combination twenty-eight subjects were randomly selected and assigned to cells in the design matrix. All subjects were followed to the end of grade one (that is over three years).

The California Test of Mental Maturity (CTMM) and Deutch index of socio-economic status were used for the initial survey to determine school composition. The Stanford-Binet Intelligence Test (SB), Deutch, and locally administered health, vision, and hearing tests were used to determine CD. The Stanford-Binet (SB), six subtests of the California Test of Personality (CTP), six subtests of the Metropolitan Readiness Test (MRT), and four subtests of the Metropolitan Achievement Test (MAT) were used as dependent variable measures on IQ, personality, readiness, and achievement respectively. The IQ was measured at the end of each year, personality at the end of years two and three, readiness at the end of year two, and achievement at the end of year three.

A four-way factorial design with repeated measures and subjects nested in the factors was used. Factor 1 was program (G) with levels as follows (each group consists of two teachers):

G1 = Piagetian curriculum (4 teachers)
G2 = Language curriculum (4 teachers)
G3 = Unit-Based curriculum (4 teachers)
G4 = Present Practices (remained at home)

Factor 2 is race with two levels, and factor 3 is sex with two levels. The fourth factor was time. For purposes of analysis, three times were used: T1 = initiation of Headstart treatments, T2 = end of Headstart/beginning of first grade, and T3 = the end of the first grade.

Results. There was an initial disparity in IQ with whites scoring significantly higher than blacks in all groups at time one. At time two and time three there was no significant difference between races within any of the treatments. At times two and three females scored significantly higher than males. At time two group one (Piagetian) scored higher than present practice; however, there was no significant difference at time three.

CTP: No significant differences were observed on treatment-related factors.

MRT: At time two, females scored significantly higher than males on number readiness. On sentence readiness, whites scored significantly higher than blacks. For males, all three experimental programs produced higher scores than present practice, but only the Piagetian and unit-based scores were significantly higher. This trend was more marked for black males than for white males, but the racial difference was not significant.

There was no significant overall difference by race, but whites were significantly higher than blacks in present practice.

MAT: The MAT was given at time three only. Females scored significantly higher than males on word knowledge, word discrimination, and math. Whites scored significantly higher than blacks on reading and math. For math, there was a treatment by sex interaction. The present practice and Piagetian groups were equally effective for girls, both being significantly better than the other two. For boys, the Piagetian and unit-based groups were both significantly better than the language and present practice groups.

Conclusions. The initial disparity in IQ by race appears eliminated by time two. However, since the racial differences in IQ were reduced even in present practice, there was no reason to attribute this elimination to Headstart intervention.

The significance of the IQ differences between males and females at the end of two and three years was probably attributable to the recognized earlier maturation of females, and not to any intervention.

There was no support for the notion that these intervention programs were especially beneficial for black minorities. In fact, the racial differences appear to have been washed out by SES and sex. The findings for the readiness tests appeared to favor the use of Piagetian and unit-based males, but not for females. However, any advantage that may accrue here did not carry over to achievement in grade one.

Although the findings do not establish the cost effectiveness of these intervention programs, the evaluation team recommended that the Fair City Board continue the Piagetian and unit-based programs again next year, but that the language program be dropped.

B. Summary of Evaluative Comments: RFP004

1. Like PEP, the HST program generated a great deal of emotional commitment. Similar criticisms of the limited nature of the treatments and the inability of trainees to define their own tests were received. Again, a consensus definition of the problem and a clear focus on the process was used to alleviate the contention where it was undesirable -- however, many early childhood instructors believed that the trainee's disagreements about the meanings of various variable scores and methods of combining them were the HST problems most valuable characteristic.

2. There were some comments about the "built-in" racism of the problem -- particularly from black students -- but these objections disappeared when it was discovered that racial differences in the FEHR data tended to disappear when the effects of sex and socio-economic status were held constant. (See the results of the illustrative study, above.)

3. Many of the trainee projects in HST yield no significant differences. Some game managers/instructors felt this was discouraging. (Note: Instructors who desire to produce large and significant treatment differences may do so with the new multiplier option described in the last chapter.)

V RFP005: READING ASSESSMENT PROBLEM (READ)

As mentioned previously, the initial problem package for READ turned out to have a serious bug in the criterion referenced variables. Consequently, the entire package was reprogrammed. The information contained in this summary is based on a report from a group of senior students in the research training program who were asked to complete a project on the READ problem in order to validate the revised version. Since one of the three team members had a considerable amout of experience in reading assessment, this team was specifically chosen to evaluate the "believeability" of the data.

The summary below is an abridged version of the actual study report.

A. Illustrative Report.

Problem

The specific problem to be attacked by this project was to determine whether the differences among the three treatments available in the READ problem -- (1) present practice (pp), (2) linguistic reading method (LRM), and (3) total language arts approach (TLA) -- were consistent with the relationships built into the problem package. These are outlined under the conceptual framework heading.

Review of the Literature

No review of the literature was assigned for this study.

Conceptual Framework (Rationale)

The basic assumptions underlying the theoretical structure of the READ problem is based on the assumption that some students have learning styles best suited to specific phonomic practice (the LSM program), and others have styles best suited to an integrated holistic approach (TLA). Since the present practice (pp) is eclectic it contains elements of both approaches. Thus one would expect pp to produce greater overall learning than either of the other programs. Overall learning in this context is defined as the percentage of students at criterion averaged over all competencies. Assuming that the students are approximately equally distributed between learning styles, one would expect no differences between the LSM and TLA treatments in overall effect. However, one would expect differences in the patterns of competencies, with each program producing best results in the competency items most directly related to the training technique concerned. Thus, LSM would be expected to produce a larger proportion of students at criterion on competencies concerned with phonemic skills, and the meanings of individual words. TLA should produce more students at criterion on competencies concerned with comprehension of sentences and paragraphs.

Subjects

The sample to be used in this study consists of three intact classrooms drawn from different schools systematically so as to represent the entire range of socio-economic status in Fair City.

Instruments

Twenty eight variables were selected from the 172 variables available in this problem. These consist of seven standardized tests and twenty one criterion referenced tests which are representative of the total set of competencies to be developed by the reading program. The standardized tests consist of both the grade 1 and 3 forms of the SAT study, SAT word, and SAT paragraph, plus the Gates Advanced Primary Test of paragraph meaning, which has only a grade 3 form. The study and paragraph tests ought to favor TLA.

Six of the criterion tests are concerned with phonemics and other linguistic skills: which should favor LSM tests 79, 81, 82, 94, 101, and 105. Seven of the tests are concerned with intergration and inference skills which should favor TLA: tests 120, 122, 129, 130, 131, 134, and 135. The remaining eight criterion tests (60, 61, 63, 92, 110, 116, 127, and 128) are concerned with general skills which should not favor either LSM or TLA. The specific competencies assessed by each of these tests are identified in the RFP package.

Design

Because of the problem validation purpose of this project, the research team took advantage of the FEHR data tenerators capacity to use the same ninety subjects in each of the three treatments, setting them "back to zero" at the beginning of each new treatment. This was done to ensure that the true treatment differences could be isolated. However, the analytic procedures will not take statistical advantage of the high correlation between subjects.

Analysis

Each of the standardized tests will be subjected to a one-way ANOVA with two subsequent planned comparisons: PP vs (LSM & TLA), and LSM vs TLA. Follwoing these ANOVAS, a single test of the probability of obtaining the observed pattern of t-test results for each comparison will be computed.

The criterion-referenced scores will be combined into three composite variables: (1) linguistic skills, and (3) overall skills. The observational unit in this case is the proportion of skills mastered: these will be analyzed in the same way as the standardized scores, above.

Results

It is important to note that the significance tests reported in this section were computed from an analysis of variance for random independent groups rather than using a repeated measures error. This was done to keep the statistical power comparable to that available in a conventional experiment.

The results of the planned orthogonal comparisons for the standardized tests appear in tabel 1. It was observed that the direction of the differences favored PP over LSM & TLA for all variables. The probability of obtaining this particular pattern of t-tests by chance was less than .001. Although four of the LSM vs TLA comparisons yielded differences favoring TLA and only one favored LSM, none of the t-tests for individual comparisons was significant, and the overall pattern was not significant.

The mean percentages at criterion on the linguistic skill composite variable for treatments PP, LSM and TLA were 79, 71, and 72 respectively. The corresponding percentages at the end of grade three were 98, 97, and 98. A conservative estimate of the standard error of the difference in percentages for these data can be computed from the variance of class means within treatments. This yields values ranging from 3.5 to 4.6 for the standard error of the difference. Thus differences less than 7 may be considered NOT significant, those equal to or greater than 9 definately significant and those between 7 and 9 as marginally significant.

Using these criteria PP was marginally better than LSM and TLA at the grade one level, with no significant difference between LSM and TLA. But by the end of grade three all differences had disappeared: the three treatments all had produced 97 - 98% criterion attainment.

The percentage at criterion for the integration skill composite variable for PP, LSM, and TLA respectively were 47, 39, and 37 at the end of grade one and 94, 85, and 87 at the end of grade three. The advantage of PP over the experimental combination (LSM & TLA) was significant at the end of grade one and still marginally significant at the end of grade three. Again there was no significant difference between LSM and TLA in either grade.

A similar pattern of results emerged for the composite overall skills. In order of treatments PP, LSM and TLA the resulting percentages were 61, 58, and 52 at the end of grade one and 94, 89 and 88 at the end of grade three. This was interpreted as a marginally significant difference favoring PP over LSM and TLA at grade one and no significant differences at the grade three level. These results are summarized in figure 1.

Conclusions

The pattern of results support the hypothesized superiority of the PP treatment overall. However, the differential patterns of effectiveness for LSM and TLA failed to materialize.

B. Summary Of Evaluative Comments: RFP005

1. This is the second version of READ to be tested. The first version was considered unacceptable because the pattern of scores it yielded were unbelievable. This completely reprogrammed version has only been used by one team.
2. The recent rational emphasis on the "right to read" programs and the current movement to state assessment of reading has created a strong interest in the READ problem.
3. Students like the problem's emphasis on criterion-referenced scores.
4. As with PEP and HST, there was some frustration expressed with their (trainees) inability to administer tests other than those made available -- many wished to administer tests of their own. However, this did not occur in the second session, where the process orientation was emphasized.
5. Trainees experienced great difficulty in establishing criteria of success which permitted them to compare treatments.
6. The hierarchical structure (i.e., the order in which various skills were learned) was not criticised. This was considered very encouraging, since it was lack of this hierarchy which had caused the first version of the problem to be unacceptable.

*7. The overall success ratio appears to be too high to be realistic. It was suggested that these be revised downwards.

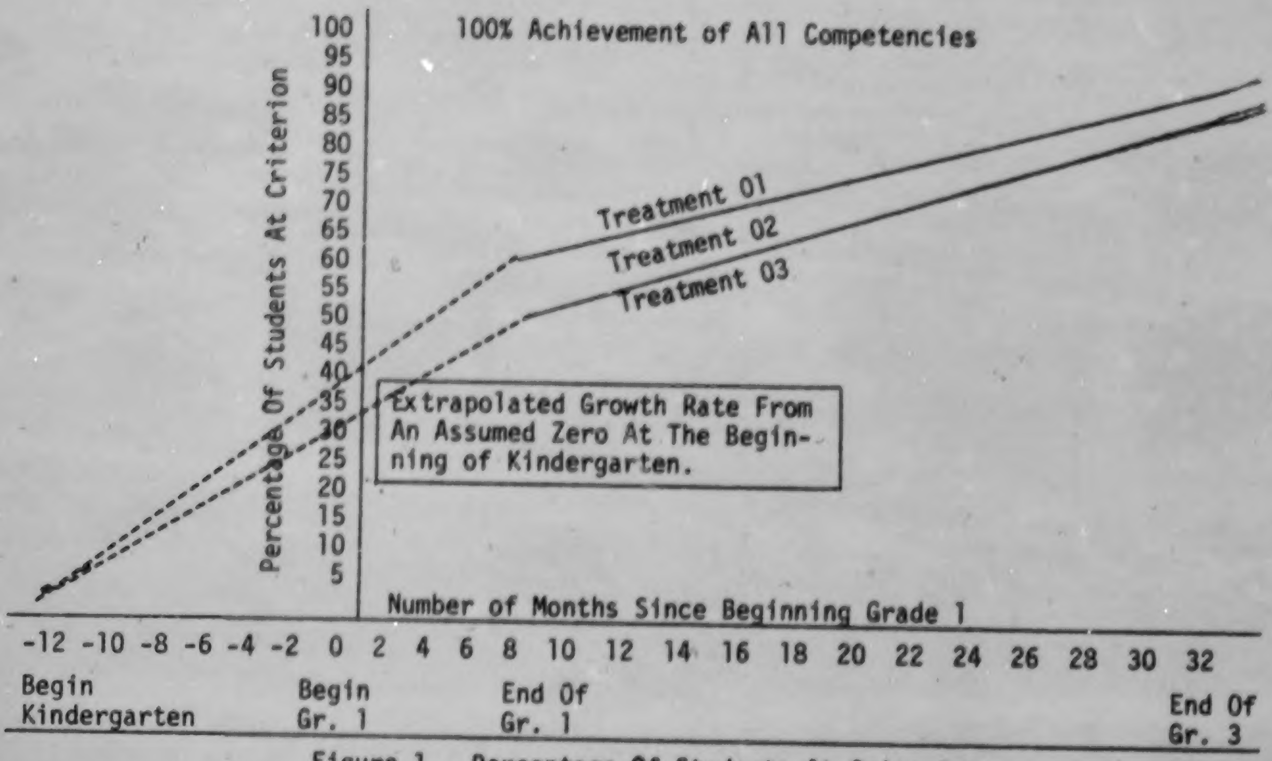
*8. There ought to be a clear and unambiguous shift in the pattern of success on criterion-referenced variables between the LSM and TLA programs. The program needs revision to accomplish this.

---

* The problem package is currently being revised to implement these suggestions.

TABLE 1. MEANS AND ORTHOGONAL CONTRASTS

TEST: GRADE:	S.A.T. ONE	STUDY THREE	S.A.T. ONE	WORD THREE	S.A.T. ONE	PARA THREE	GATES THREE
PP Means	36.47	55.50	19.93	30.80	18.60	52.80	25.57
LSM Means	34.50	49.13	19.03	28.20	16.63	47.47	23.37
TLA Means	34.50	50.00	18.70	28.20	17.27	47.80	23.97
ANOVA SUMMARY							
MS between groups	116.43	1073.76	36.48	202.80	90.89	655.32	116.40
MS within groups	82.93	158.93	44.65	41.65	40.23	102.07	17.84
F(2,270)	1.40	6.76	0.82	4.87	2.26	6.42	6.52
CONTRAST T. RATIOS							
PP vs. (LSM & TLA)	1.70	3.71	1.26	3.17	2.05	3.64	3.54
TLA vs. LSM	0.00	.47	-0.34	0.00	0.60	0.22	0.97



Extrapolated Growth Rate From An Assumed Zero At The Beginning of Kindergarten.

Figure 1. Percentage Of Students At Criterion Averaged Over All Competencies.

VI. RFP006: VALIDATION OF A NEW TEACHER QUESTIONNAIRE (TQUES)

The information contained in this summary is based on a report completed by FEHR trainees 212, 241, and 246. The problem with which they were concerned was set in the School of Education of Utopia University in Fair City, U. S. A. The school has for many years routinely administered a questionnaire entitled "Student Course Evaluation" at the conclusion of each semester. However, a number of faculty have complained, over the years, that the instrument provided little information to help them plan needed changes in their course. Recently, a student-faculty committee has developed a new questionnaire which they claimed would provide measures of the dimensions of classroom performance over which an instructor has control. This evaluation team consisted of three students in a second semester research course. They were assigned the task of validating one of the dimensions defined in the RFP document. The final report for their project is summarized below.

A. Illustrative Report

Problem. The researchers wished to determine whether the feedback of summarized scores on the personal factor of the new questionnaire provided instructors with information which was not contained in the summarized scores from the old questionnaire. The personal factor, as defined by the RFP, consisted of students' ratings of: the adequacy of the individual help provided by the instructor, the degree of instructor concern for the progress of individual students, the amount of effort he or she (the student) had put forth in the course, and the work load of the course relative to other courses.

Hypotheses. The researchers hypothesized that feedback should have a generally positive effect on ratings of the "individual help" and "concern for student" items. However, this effect should be much greater when there was a high need for help and a low availability than when there was a low need and a high availability.

Method. The subjects for the evaluation were eight intact classes (189 subjects). Four of the classes were relatively small laboratory-type classes, and four were large lecture-type classes. It was assumed that the instructors for laboratory classes were more available for individual help than were instructors for the lecture classes. Two classes of each type were assigned to an experimental condition in which instructors at the beginning of the current semester were given feedback on the personal factors items of the new questionnaire as well as the usual information from the old questionnaire. The remaining classes (two of each type) were assigned to a control condition which received only old questionnaire information.

All eight groups were administered both questionnaires at the end of the semester. Subsequently, all students were classified as "low need" or "high need" depending on their average score on the effort and difficulty items on the new questionnaire: Anyone with an average of 3.5 or higher was considered high need, those with averages below 3.5 were low need.

Results. The dependent variable of interest was the average of the "individual help" and "concern for student" items. These average scores were analyzed in a three-way factorial analysis of variance (treatment x availability x need). The results of this analysis were:

1. Ratings in the experimental groups were significantly higher than those of the controls.
2. The high availability group had significantly lower ratings than those in the low availability group.
3. There was a significant treatment by availability interaction. The experimental vs. control gain was larger for the high availability subjects (labs) than for the low availability classes (lectures).

4. There was a significant treatment by need interaction. The experimental vs. control gain was greater for low need than for high need classes.

5. There was a significant three-way interaction. In the control condition, the low need groups gave lower ratings in both high and low availability settings. In the experimental condition, low need groups gave somewhat lower ratings than high need in the low availability settings, but gave considerably higher ratings in the high availability setting.

Conclusions. It was concluded that since feedback of the personal factors information from the new questionnaire did produce a difference in perceived behavior, the new questionnaire items must yield information not contained in the old questionnaire. However, the failure to discover a need x availability interaction raised the question of whether the lecture/laboratory distinction was actually an availability variable. Further research to clarify this point was suggested.

B. Summary of Evaluative Comments: RFP006

1. The TQUES problem was considerably more didactic than the first four in that a criterion of success was spelled out by the theory supplied in the problem. Consequently, there was very little outside reading stimulated by this problem.

2. Many trainees felt that the theory upon behind the questionnaire to be evaluated was somewhat weak.

3. Most trainees felt that the absence of the capacity to question respondents about their interpretation of (questionnaire) items was a real weakness.

4. Nevertheless, virtually all trainees reported that TQUES gave them valuable insights into the difficulties of questionnaire validation.

VII. RFP007: REMEDIAL MATH FOR ADULTS (RMA)

The information in this summary is based on a report by FEHR trainees 40 and 68. Their assignment was to make an assessment of the effectiveness of the free "remedial math for adults" program offered by the Fair City Community College. The course consists of a series of programmed lessons which each student works at his own rate. There are no formal classes, but each student is assigned to an instructor who provides immediate feedback on the adequacy of each lesson and one-to-one tutorial help during interviews scheduled at the student's request. Only the one treatment (RMA) is available in the problem -- no control group can be specified. This evaluation team consisted of two members of a beginning research class. At the time of the study, they had covered no statistics beyond t tests and a simple one-way analysis of variance. Their report is summarized below.

A. Illustrative Report

Problem. The researchers were interested in whether the RMA program produced significant growth in mathematics achievement as measured by the computation and concepts subtests of the Stanford Achievement Tests for mathematics (grades 7-9), and whether there was a concurrent reduction in the perceived difficulty of mathematics problems.

Hypothesis. It was hypothesized that all the dependent variables should be positively affected by the treatments, but that concepts -- which were directly targeted by the program -- ought to be more affected than computation, which was only incidentally taught. Attitude was expected to improve significantly.

Method. A sample of 120 subjects were randomly drawn from the 216 who qualified for the course. Each of these subjects was pretested on the SAT computation and SAT math concepts tests for grades 7 to 9, and asked to indicate the difficulty of mathematics problems for them on a one to five scale. When they had completed all the lessons, each subject was post-tested on the same three measures. T tests for matched samples were run on each variable to assess the significance of the obtained difference in means. Subsequently the mean gain scores for the two SAT tests were converted to normal deviates by dividing the obtained difference by the published standard deviation for the norm group on that test. The difference between the two normal deviate gains was then entered into a t test for independent samples.

Results. The obtained results appear in the table below. It was observed that there was a significant gain in concepts and in attitude, but not in computation. Also, the gain in concepts was found to be significantly greater than the gain in computation.

Test	Means			Paired t-test	Norm Dev. Gain	Indep. t-test
	Pre	Post	Gain			
Computation	13.80	14.39	.59	1.51	.0667	8.69**
Concepts	16.92	26.68	9.76	13.11**	1.1888	——
Attitude	2.08	2.64	.56	6.43	——	——

** $p < .01$ * $p < .05$

Conclusion. While it was recognized that the fact that the same group was used to obtain the computation and concept scores, it was argued that the use of an independent t test was actually a conservative test. Since concepts gained significantly more than computation -- which should have been subject to identical history effects -- it was concluded that the RMA program produced significant growth in mathematical concepts. Although no control comparison was possible, it was concluded that there was probably some significant reduction in the perceived difficulty of mathematics. (Note. The attitude scale ran from 1 = difficult to 4 = fairly easy.)

B. Summary of Evaluative Comments: RFP007

1. The RMA (remedial mathematics for adults) problem is perhaps the most restrictive of the eight problems. Since there is only one treatment, no control group is possible. Nevertheless, most users felt that this was valuable because of the prevalence of real-life situations of a similar nature. Experience with this problem, many felt, emphasized the inadequacies of one-shot single-group studies.
2. Several trainees expressed "amazement" that variables which theoretically ought nought to be affected by the treatment really remained constant -- they had not believed the simulation to be that thorough. (Note. In some cases, as our illustration shows, such variables as these were used as "controls".)
3. Several students in higher education programs felt that this problem was the closest approximation to their real life situations.

VIII. RFP008: BUSING TO ACHIEVE INTEGRATION (BUS)

The information contained in this summary is based on a report completed by FEHR trainees 22, 50, and 57. Their general assignment was to evaluate the effects of the busing program which the Fair City School Board has recently voluntarily implemented in an attempt to overcome de facto school segregation attributable to the existing housing pattern. Since the researchers were not called in until just after school opened with the new busing system, there was no opportunity to obtain pre-measures or to organize a controlled experiment. With the exception of the meagre information available from the files, the researchers must rely on measures taken after the project was begun. Funds were available for a three year longitudinal study of two grade levels: the present grades one and four.

The research team consists of three members of a special class in program evaluation. All class members have had at least two courses in research design and data analysis prior to completing this project. The illustrative summary which follows is a simplified abstract from their comprehensive final report.

A. Illustrative Report

Problem. After an extensive review of the findings of the Coleman report supplemented by a variety of additional research, the team decided to focus on the effect of the busing program on the relative performance of black and white students on academic performance in reading and mathematics and on their attitude towards school.

Hypotheses. On the basis of the Coleman findings, the researchers hypothesized that:

1. When academic performance was measured in standard scores based on the appropriate national norms for the Gates reading comprehension and SAT arithmetic concepts tests, black students would improve their relative position over the three year evaluation term but white students would maintain about the same relative position.
2. Attitude towards school, expressed as ascending scores on a dislike to like continuum would increase significantly for black students and exhibit no change for white students.
3. The above relations would hold at both the grade 1 and grade 4 levels.

Method. A stratified random sample of 800 subjects was selected with approximately equal representation of each race at both grade levels. To maximize the possible effect of busing, students were chosen only from attendance zones A and C which were maximally affected by the busing decision.

The SAT and Gates tests, and an attitude questionnaire were administered at the beginning of the project and again at the end of the three year evaluation period. Scores on the two standardized tests were in each case transformed to Z scores using the national norms appropriate for their grade. T tests for correlated samples were then conducted to test the null form of the above hypotheses. [Note. The research team had been informed by the game manager that there would be no attrition during their experiment as this part of the FEHR program had been switched off for this session.]

Results.

1. At the grade one level, the mean Z scores for black subjects were significantly higher at the end of the evaluation on both the Gates and SAT tests.
2. At the grade four level, the mean Z scores for black subjects were significantly higher for the Gates test but not for the SAT (although even this difference was in the hypothesized direction).
3. There were no significant shifts in mean Z scores for whites at either grade level.
4. Both races showed increasingly positive mean attitude scores over time at both grade levels, but none of these differences reached significance at the .05 level.

Conclusion. The results were considered supportive of the hypotheses in all cases. It was concluded that the busing project was a success in terms of the selected criteria.

B. Summary of Evaluative Comments: RFP008

1. Many trainees expressed interest in the substantive ideas behind the Fair City Busing plan. Such comments as "This is the best plan I've seen." and "I like this plan." were common.
2. One game manager/instructor expressed surprise at the "consistent pattern of favorable results" and wondered whether it was realistic.
3. Two instructors questionned the desirability of constraining the problem to post hoc studies. Is it good

research training to require trainees to do such studies. However, most others felt that many real-life factors resulted in similar constraints, and that trainees ought to have experience working within such constraints.

4. These findings agree with the most hopeful of real-life studies. Is this good training for would-be researchers?
5. In view of the recent court decisions against busing, some reduction of interest in the problem has been experienced.

INTERNAL VALIDITY OF FEHR-PRACTICUM MODEL

It is the position of the authors that the only evidence for the validity of any simulated problem which is necessary is a demonstration that it, in fact, stimulates in its users the type of behavior which it was designed to produce. The projects summarized in this chapter provide concrete evidence that each of the eight FEHR problems is capable of motivating trainees to the kinds of problem solving behavior typical of the research/evaluation task. We conclude that in this sense all eight problems are valid simulations.

The successful simulation of eight different problems demonstrates the internal validity of the underlying FEHR-PRACTICUM model. In each case the generated data was considered believable and realistic, and participants frequently reported feeling a sense of urgency and an emotional involvement similar to that experienced in the real-life situation. We conclude that the model is sufficiently flexible and adaptable to simulate a wide variety of different problems. The more important question of the pedagogical effectiveness of the FEHR experience is attacked in the next chapter.

CHAPTER 4

SUMMATIVE EVALUATION

The summative evaluation presented here varies in only a few minor details from the plan presented on pages 5 to 11 of the quarterly progress report, dated October 20, 1972. Most of the departures from the plan are attributable to the decision (in late October 1972) to drop the restricted, structured, unstructured problem designations in favor of a design which permitted instructors to define their own structures and/or restrictions. There were now a potentially infinite number of problem "levels" -- essentially a continuum of complexity/difficulty. The evaluation plan had to be expanded to accommodate this increased range. This expansion was possible because of the decision to discontinue development and evaluation of the performance contracting problems nine and ten. The rationale and philosophical justification for these decisions was provided in a previous section: they are mentioned here merely to help delimit the task.

Purpose

The purpose of the summative evaluation was to assess the degree to which FEHR-PRACTICUM achieved its general educational objectives. The system was developed with eight major objectives in mind. Broadly stated, these objectives were:

General Achievement Objectives

Objective 1. To improve achievement in the content area traditionally associated with research/evaluation training: measurement, experimental design, statistics, data analysis by canned computer programs, and the like.

Objective 2. To develop the ability to write proposals and final reports which are explicit, operational, well organized, and sufficiently comprehensive to permit replication.

Objective 3. To encourage effective field studies; viz., those which feature:

(a) designs which contain a control group and which permit valid contrasts on each of the critical study dimensions.

(b) multiple dependent variables. (It is assumed that in most practical situations the use of a single dependent variable is a gross oversimplification leading to costly errors of omissions.)

(c) an attempt to assess the cost effectiveness of both the programs being evaluated and the evaluation procedure per se.

General Attitude Objectives

Objective 4. To increase interest in research and research methods generally.

Objective 5. To increase the perceived relevance of both the methods and practice of research and evaluation.

Objective 6. To foster a positive attitude towards the computer.

Objective 7. To foster a positive attitude towards teamwork.

Summary Objective

Objective 8. To provide instructors with an adaptable research evaluation practicum which can facilitate a wide variety of instructional purposes.

Critical Comparisons

The design of the summative evaluation was dictated by the critical comparisons implicit in the objectives. Within each objective, the following four comparisons were considered critical:

Contrast 1. The first, and most important, critical comparison is the usual experimental versus control condition. Ideally, the control for an "independent FEHR course" would be a course in program evaluation methodology which did not use the practicum course. However, no such course is presently offered at any of the available sites. Only the traditional courses in research design, statistics, and measurement were available. (It is apparently assumed that the transfer to classroom-based field studies and quasi-experimental design will occur automatically -- an assumption which we question.) For these reasons no direct control was possible for the "workshop" condition.

Two types of controls were available for the "integrated FEHR course". The author was scheduled to teach two sections of a research design and data analysis course sequence during the evaluation interval. It was feasible to develop an experimental "integrated FEHR" condition for one section and use the other section as a control. Two other statistical courses were available as outside controls.

In addition to the above controls for the integrated research training class, it was desirable to have a set of subjects with no training or experience in research/evaluation to provide a comparative base for the attitudenal dimensions. Students from a core course in educational philosophy were available. Since all graduate students were required to take the course, the class was deemed an adequate control for this purpose.

Contrast 2. The second critical contrast concerns the relationship between effectiveness and amount of experience with FEHR. Experience in this context is increased by increasing either the complexity level at which a problem is attacked or the number of problems "solved", or both. This is really just an extension of contrast one, since the control condition may be defined as zero experience with FEHR.

Contrast 3. The third critical contrast concerns the problem content: What happens to the effectiveness of the game as we move from problem to problem? Are all problems equally effective?

Contrast 4. The fourth contrast is concerned with the interrelationship between FEHR-PRACTICUM and existing research evaluation courses. FEHR-PRACTICUM was conceived as a vehicle for upgrading the program evaluation skills rather than a self-contained training package. Consequently, it is critical to provide a comparison of the effects of integrating the package into a formal pre-structured course (or program) versus using the PRACTICUM experience as the "syllabus" and providing whatever consultation (teaching) is necessary for the player to "solve" the problem. For convenience, we shall refer to the latter usage as a "FEHR workshop" in subsequent discussions.

Organization

The remainder of this chapter is organized in two sections. A brief preamble at the beginning of each section describes its contents and structure. Section I contains a narrative description of the summative evaluation process. Section II is devoted to a detailed technical presentation of the empirical evidence. However, the summarization, integration and interpretation of the findings with respect to the educational objectives is not included in this section. For the convenience of the reader, this material is presented separately in chapter 5.

SECTION I. NARRATIVE DESCRIPTION

This section is intended to provide the reader with an overview of the entire project before proceeding with the specific details of data analysis and interpretation. The discussion is organized under two main headings: subjects and instrumentation. Under the subjects heading we provide detailed descriptions of the various settings in which the trials occurred, and operational definitions of the major independent variables of interest to the evaluation: viz., problem content, amount of FEHR exposure, type of class, and degree of integration between the regular class contents and the FEHR-PRACTICUM project. Under the instrumentation heading we provide a detailed description of each evaluation instrument used, and the process by which it was developed and validated. Hopefully, this procedure will permit the presentation of the empirical data to be shorter and better articulated than would otherwise have been the case.

SUBJECTS

The summative evaluation of FEHR-PRACTICUM involved 358 subjects from 20 different education classes conducted during the 1972 and 1973 calendar years. The majority of these (15 classes and 306 subjects) were regular course offerings at The University of Michigan. The remaining 52 subjects were distributed among experimental courses offered by five different institutions: Flint Junior College, Indiana University, Michigan State University, Ohio State University, and Western Michigan University. Since three of the University of

Michigan classes (50 subjects) were used as controls, a total of 308 FEHR-PRACTICUM experiences were evaluated (256 at Michigan and 52 elsewhere). However, the effective sample of experimental subjects was only 215 because 93 subjects appeared twice. Double appearances occurred when students enrolled in both terms of a two-semester sequence. These repeated administrations were considered even more valuable than an equivalent number of new subjects. In addition to the longitudinal information provided by these cases, each separate appearance contributed unique information because the instructional purposes, the FEHR-PRACTICUM problem used, and the administrative procedure were different in the first and second semesters.

Originally, we had planned to collect a uniform set of data from each field trial. This was a practical plan when the system consisted of ten rather finite problems. However, the current FEHR-PRACTICUM system permits each user (instructor) not only to choose which of the eight problems he will use but also to adapt the practicum to the needs of his students by assigning only those tasks which are directly related to the instructional objectives of the session. Literally hundreds of "assigned problems" with differing levels of complexity and difficulty (attained through different task combinations) are possible within the general framework of each FEHR-PRACTICUM problem. Detailed descriptions of the more important task combinations and comprehensive instructions for their use are provided in the FEHR-PRACTICUM Game Managers Manual. It follows that each combination would imply instructional objectives with different patterns of emphasis. Evaluating the effect of FEHR-PRACTICUM in terms of these differing purposes obviously required different data bases or different interpretations of the same data, or both. Because of this interrelationship it is necessary to discuss the particular instructional objectives to be included before describing the data base *per se*.

It is obvious from the above discussion that the number of possible problem variations precluded an evaluation of each FEHR problem for all instructional purposes. Even an attempt to evaluate each problem with a set of say four typical instructional uses would have required 28 subject groups for each replication. There was also the fact that an adequate evaluation in many problems required a comprehensive and detailed knowledge of a specialized content area. For example, the perceptual education problem (PEP) was designed for clients who either already had a broad knowledge of theory and practice in the psychology of learning disabilities or were willing to spend the time and effort to develop it. Similarly, the Headstart problem required interest and knowledge in the field of early childhood education. Clients with these interests were not available in sufficient numbers to permit an evaluation of a wide range of complexity and difficulty levels for assigned problems. For these reasons, the strategy adopted was to evaluate the flexibility/adaptability of the FEHR system using one "standard" problem, then to field test each of remaining problems at a complexity level which ensured that each component task was involved. The remedial arithmetic problem (REMAR) was chosen as the standard problem because it seemed reasonable to assume that most prospective clients had sufficient experience and expertise in computation to develop an adequate evaluation rationale. Although the use of a FEHR-PRACTICUM problem to motivate the development of content expertise was considered a legitimate function for the game, we chose not to evaluate this usage because of the prohibitive amounts of time involved.

Class Settings and Instructional Objectives

The seventeen classes in which FEHR-PRACTICUM was field tested can be divided into two broad groups. According to our evaluation plan, the first nine classes were to be used for field testing the eight FEHR problems (RFP packages), and the last eight classes were to be used for field testing the flexibility/adaptability of the FEHR problem model using only the standard REMAR problem. However, in practice the distinction between the two groups was blurred by differences in the innate complexity and difficulty of the problems themselves and wide variations in the expectations and standards of the instructors and game managers from site to site. In addition there were wide variations in the entry skills of the participating subjects (students) from class to class. For these reasons the original dichotomy was dropped in favor of a three dimensional classification scheme.

The three dimensions of interest are: (1) the problem content area (or areas) used, (2) the degree of exposure to FEHR (number of projects, and the tasks assigned in each), and (3) the type of class. A brief description of the categories within each dimension is provided before describing the sampling pattern per se.

(1) Problem Content: There were eight problems to be evaluated in the FEHR-PRACTICUM model. A list of the titles is provided below. A detailed description was given in the previous chapters.

- i. Project PEP: Perceptual Education Problem (RFP001).
- ii. Project REMAR: Remedial Arithmetic (RFP002). This is the standard problem described to be used for the first implementation at a new site.
- iii. Project EXTSY: Extended School Year (RFP003).
- iv. Project HEADSTART: Early Childhood Education (RFP004).
- v. Project READ: Reading Assessment Project (RFP005).
- vi. Project TQUEST: Validation of a Teacher Rating Questionnaire (RFP006).
- vii. Project RMA: Remedial Math for Adults (RFP007).
- viii. Project BUS: Busing to Achieve Integration (RFP008).

(2) Exposure to FEHR. The amount of exposure to FEHR depends on both the number of FEHR projects a subject participates in, and the complexity of each project. The complexity dimension is operationally defined in terms of the specific tasks which were assigned in a given class. The items on the checklist of practicum tasks included in each problem RFP packet can be subdivided into eight main categories. Listed in order of occurrence with a section keyword underlined these are: (a) introduction and problem definition, (b) review of the related literature, (c) conceptual framework or theory, (d) method, (e) plan for analysis of data, (f) personnel responsibilities, logistics, and budget, (g) results of the analysis, and (h) the educational interpretation and a recommended decision.

In general, the eight task categories were assigned in five main patterns. These are listed below in (approximate) order of increasing complexity. Pattern D was not listed as a FEHR-PRACTICUM assignment in this evaluation but was included to illustrate the relative position of the dissertation proposals used as a comparative criterion in the evaluation of FEHR proposals.

Pattern A: A restricted statistical study only. Contains: Problem, method, analysis and results.

Pattern B: An experimental report, with the review. Contains: Problem, theory, method, analysis, results and interpretation.

Pattern C: Both a proposal and a report. Contains: Problem, theory, method, analysis, logistics, results and interpretation.

Pattern D: A full proposal. Contains: Problem, review, theory, method, logistics, analysis.

Pattern E: Both a full proposal and a full report. Contains: Problem, review, theory, method, logistics, analysis, results and interpretation.

Anyone of the patterns outlined above could be completed with varying degrees of sophistication. For example, reading achievement in problem five could be defined as the total score on a single standardized reading test, or it could be defined as a pattern of scores on a series of sub-tests. Obviously the latter definition requires a greater understanding of psychological theory, measurement constructs, and data analysis techniques than the former. This we called the intensity dimension. For the sake of simplicity, all practicum sessions were classified as either intensive or non-intensive. An intensive session required subjects to develop a detailed theoretical structure (usually multivariate) which was comprehensive, internally consistent, and clearly related to previous evidence in the field. The complexity and intensity dimensions were combined to make the exposure factor according to the following rules.

Rule 1. The higher the pattern level, the greater the exposure. That is, pattern E > pattern D > pattern C > pattern B > pattern A.

Rule 2. Two experiences at any pattern level represent more exposure than one experience at a higher pattern level provided the experiences are with different problem contents.

Rule 3. One intensive experience represents more exposure than two non-intensive experiences.

Rule 4. A non-intensive experience followed by an intensive experience represents more exposure than a single intensive exposure, but less exposure than two intensive experiences. Two intensive experiences represented the maximum possible exposure available in this study.

(3) Type of Class. Seven different courses were represented in the 17 which used FEHR-PRACTICUM. Each of these carried graduate credit in the general area of educational research, but their clientele and purposes differed considerably. However, it is convenient to group them in three course types: general research methods, research methods for specialized content areas, and in-service workshops for practicing educators. The particular courses in each classification follow:

(i) General Research Methods. These courses were attended (usually on a required basis) by students from a variety of graduate programs in education. The courses in this category were:

(a) Education 882 at Michigan State University, taught by Mr. George Sargent in collaboration with Professors Norman Bell and Allan Abedor. The students had already completed elementary statistics and were studying research design and analysis of variance, but typically had little previous mathematics or research training. Teams of three were assigned a REMAR project at the pattern B level of complexity. A somewhat structured FEHR-PRACTICUM (i.e., one dealing with 4-6 variables) was used as the core content of the course, with lectures, seminars, and self-study materials paralleled the problem-solving process. Course material was parallel with the practicum but not completely integrated into the curriculum. The class was considered non-intensive.

(b) Education 785: Introduction to Inquiry at Ohio State University, taught by Professor William Loadman. This course was very similar to the course at Michigan State University except that the pattern C level of complexity was used in order to emphasize budgeting and negotiations. In addition to formal written budgets including cost-effectiveness assessments, Professor Loadman required each team to meet with him to negotiate their project funding. Again, all students used a fairly structured REMAR problem. Course content was not integrated. The project was classified as a non-intensive experience.

(c) Education 601: Introduction to Educational Research at Western Michigan University, taught by Professor Uldous Schmidkens. This was primarily a statistics class, and a structured (univariate) REMAR problem was assigned, at the pattern A level, to a laboratory exercise to provide opportunity to practice the techniques taught in class. The course content was not integrated with the practicum. The project was classified as a non-intensive experience.

(d) Education C655 and C656, a two-semester sequence at The University of Michigan, taught by Professor LeVerne Collet, director of the FEHR-PRACTICUM project. This course sequence was of special importance to the project since it was possible to adapt the content to take optimum advantage. In addition, the existence of two separate sections enabled some experimental controls to be exercised. Eight groups (course sections) of students from these classes were used in the field trials: two sections of C655 in the fall semester 1972, two sections of C656 in the winter semester 1973, two sections of C655 in the fall semester 1973, and two sections of C656 in the winter semester 1974.

The two sections of C655 enrolled in fall 1972 were used for a formal experimental evaluation of the effects of FEHR-PRACTICUM. One section was assigned to the usual laboratory practice and the other was formed into three-man teams and required to complete a FEHR-PRACTICUM project at the pattern E level of complexity. This was a moderately restricted FEHR-PRACTICUM in which students dealt with only one dependent variable, two or three independent variables (including the treatments), and a few of the budgeting problems. The practicum was run as an independent laboratory in that none of the FEHR problems was dealt with in the lecture. However, a great deal of consultation was available during the laboratory sessions. Each team was given their choice among the seven content areas available.[1]

In the sequential C656 course in winter 1973 all students were assigned to complete a REMAR project, thus enabling us to assess the relative efficiency of early versus delayed exposure to a moderately restricted FEHR-PRACTICUM problem used as the core experience with the course designed around the PRACTICUM. It placed more emphasis on covering classical research methods than the C655 practicum. A more detailed discussion of the design and strategy of the experimental evaluation appears later in this report. The content of the course was not integrated with the project. The project was classified as a non-intensive experience.

The 1973-74 sequence of C655-C656 was used to test the notion that one cannot obtain optimum benefit from FEHR-PRACTICUM in the classical lecture and laboratory approach. To obtain optimal results, the course must be structured in a problem solving discovery mode. The pattern E level of complexity was used in C655 with the REMAR problem and the pattern B level in C656 with the TQUEST problem. In both cases the course and practicum were fully integrated. The project was considered an intensive experience. Again, a detailed explanation of this strategy appears in a later section.

---

[1] The original choice was among all eight problems, but, as mentioned previously, the READ problem experienced technical difficulties. This necessitated shifting students choosing that area to another problem.

(ii) Research Methods for Specialized Content. Only the Special Education Evaluation Practicum, a course taught by Professor Candy Garrett at Indiana University fell into this category. Students in this course were all training to become researchers, developers and teacher-educators in the area of special education. These students were required to conduct a complete project in the Perceptual Education Problem (PEP). There pattern E level of complexity was used, but the statistical aspects of the methods were downplayed somewhat and great emphasis placed on the review of research, the theory (conceptual framework) and the instrumentation. Particular attention was paid to diagnosing patterns of test scores. Consequently, it was considered an intensive experience, although this classification was marginal. The practicum and course content were parallel, but not integrated.

(iii) In-service Workshops. Three workshop-type classes were held under the course title Education C699 Program Evaluation Laboratory. The first two were held on the University of Michigan main campus at Ann Arbor. The clientele for these courses were about 75% from graduate programs in education and 25% curriculum supervisors and members of the Office of Research and Evaluation for the Ann Arbor School System. Both these groups were given their choice among the seven problems, and both required proposals and final reports at pattern E level of complexity.

The third C699 course was a true in-service course held in the Flint Junior College. It was attended by practicing administrators and administrative interns exclusively, each of whom completed a REMAR project at the pattern B level of difficulty. The aim was to develop the knowledge and skill necessary to use empirical cost-effectiveness evidence in arriving at decisions about programs. The FEHR-PRACTICUM problem was accompanied by comprehensive consultation from the instructor: here, the PRACTICUM _per se_ became the "course." However, since none of the participants had previous training in research, it was necessary to spend a good deal of time developing elementary statistical concepts. This was arbitrarily considered non-integrated because no structured content was presented. Consequently, the project was considered to be non-intensive.

In the preceding pages we provided a general description of each class participating in the evaluation, the way that FEHR was used with the class, the FEHR problem used, and the specific tasks students were assigned to complete. This information has been summarized in figure 4.1 to provide a convenient reference point for subsequent discussions.

INSTRUMENTATION

Seven formal instruments were developed to measure the degree to which subjects had achieved the eight FEHR objectives: six measures of achievement and three measures of attitude and perceived achievement. In addition, space for comments was provided on each instrument, and criticisms were solicited. However, only two instruments were administered to all subjects. The remainder were developed especially for the controlled experimental evaluation involving only a small subset of the subjects (classes 1-6). A summary of the instruments developed and the classes to which each was administered appears in figure 4.2.

None of the measures listed in figure 4.2 is considered to be a unidimensional scale. Rather, each consists of two or more conceptually independent subscales. Scales are considered conceptually independent if there is no _logical_ reason for variation in one scale to _cause_ variation in the other. Such scales _may_ yield significant statistical correlations, but these are attributed _by definition_ to common causal relationships. The evaluation _per se_ is made in terms of the subscales. Each of these scales is assumed to have primary

CLASS	N	TYPE OF CLASS	FEHR TASKS ASSIGNED	INTENSITY LEVEL	INTEGRATION LEVEL	PROBLEMS
Experimental Evaluation						
1. C655 (UM) Sec. 1, F 72	25	Gen. Res.	None	Non-intensive	Incomplete	None
2. C655 (UM) Sec. 2, F 72	22	Gen. Res.	Pattern B	Non-intensive	Incomplete	Choice 1-8
3. C655, Sec. 2 & K680, F 72	6	Gen. Res.	Pattern E	Non-intensive	Incomplete	Choice 1-8
4. C656, W 73: (a) S's from (1)	15	Gen. Res.	Pattern C	Intensive	Incomplete	Choice 1-8
(b) New S's	18	Gen. Res.	Pattern C	Intensive	Incomplete	Choice 1-8
5. C656, S's from (2), W 73	21	Gen. Res.	Pattern C	Intensive	Incomplete	Choice 1-8
6. C656, S's from (3), W 73	6	Gen. Res.	Pattern C	Intensive	Incomplete	Choice 1-8
Field Trials						
7. Mich. State U., W 73	19	Gen. Res.	Pattern B	Intensive	Incomplete	REMAR
8. Ohio State U., W 73	13	Gen. Res.	Pattern C	Non-intensive	Incomplete	REMAR
9. West. Mich. U., W 73	5	Gen. Res.	Pattern A	Non-intensive	Incomplete	REMAR
10. Indiana U., W 73	8	Spec. Ed. Res.	Pattern E	Intensive	Incomplete	PEP
11. C699 (UM), F 72	6	Prog.Eval. Wkshp.	Pattern E	Non-intensive	Incomplete	Choice 1-8
12. C699 Flint, W 73	7	Prog.Eval. Wkshp.	Pattern B	Non-intensive	Incomplete	REMAR
13. C699 (UM), W 73	6	Prog.Eval. Wkshp.	Pattern E	Intensive	Incomplete	Choice 1-8
14. C655 (UM) Sec. 1, F 73	46	Gen. Res.	Pattern E	Intensive	Complete	REMAR
15. C655 (UM) Sec. 2, F 73	30	Gen. Res.	Pattern E	Intensive	Complete	REMAR
16. C656 (UM) Sec. 1, W 74	32	Gen. Res.	Pattern C	Intensive	Complete	TQUEST
17. C656 (UM) Sec. 2, (a) W 74	19	Gen. Res.	Pattern C	Intensive	Complete	TQUEST
(b) New S's	4	Gen. Res.	Pattern C	Intensive	Complete	TQUEST
FIELD TEST TOTAL N	308					
Control Groups: Field Trials						
18. A705 (UM), W 73	30	Phil. of Educ.	None: S's had no training yet, but it was required.			
19. C607/C654, W 73	11	Theory Msrmnt.	None: S's studied class. msrmnt. & statistics elsewhere.			
20. Ph.D. Proposals (sampled randomly from Ed. Psych. files).	9	Ind. Res.	No FEHR, but proposal should be Pattern C.			
GRAND TOTAL N	358					

Figure 4.1. Summary Description of Participating Classes.

DESCRIPTION OF INSTRUMENT	CLASSES USING INSTRUMENT
Measure of Achievement	
1. First Examination: A review of basic statistical knowledge.	Exp. Eval. Classes 1-3
2. Final Examination: C655, Fall 1972, Applied Statistics.	Exp. Eval. Classes 1-3
3. FEHR-PRACTICUM rating sheet/ or proposals and reports.	All FEHR Classes 1-17 and Ph.D. proposals: Class 20
Measures of Perceived Achievement, Research Attitude	
4. Self Assessment of Research and Evaluation Skills (SARES)	Exp. Eval. Classes 1-3 and 4-6
5. Goal Assessment Questionnaire	Classes at UM and IU: 4-6, 10-17
6. ORS Questionnaire	Classes 4-17
Other Measures	
7. Written comments and criticisms were solicited at each administration of an instrument or instruments.	Classes 1-20

Figure 4.2. Summary of Instruments used in the Summative Evaluation

validity. That is, the scale is an operational definition of the characteristic being measured. Consequently, the composition rule and rationale for each subscale are of paramount importance. This information is provided in the comprehensive description of each instrument which follows. For each instrument the discussion is organized into five parts: a general description of the instrument and the process by which it was developed, the evaluation role of the instrument, the subscales derived from the instrument, and procedures for obtaining reliability estimates.

First Examination: A review of Basic Knowledge

Description. The test is comprised of 57 multiple-choice items selected from chapters 8 to 15 of Runyan and Haber (1967). The test required a broad basic knowledge of elementary descriptive statistics and simple statistical inference including t tests for either independent or correlated samples. No subscales were defined for this test. The total test score was represented by the symbol E1.

Role. The first examination was administered to the experimental evaluation only (classes 1-3) on October 16. Since the FEHR treatment did not begin until after that date, the E1 scores were used as a covariate to correct for initial differences in statistical ability among classes.

Subscales. No subscales were developed for this test.

Reliability. The split-half reliability of the entire first examination (57 items) was .83.

Final Examination: C655, Fall 1972.

Description. The instrument was administered in two parts. Part one consisted of six short answer explanation items concerning knowledge of basic statistics and six brief problems requiring the application of these ideas to simple data sets. Part two consisted of a brief description of a case study followed by questions requiring a critical evaluation of three alternative methods of analyzing the data described in the case study. All of the items in both parts had been successfully used with previous classes and a detailed scoring guide had been developed. The entire test in condensed format, appears in appendix 4A.

Role. The test was administered to the experimental evaluation groups only (classes 1-3) in two sittings. Part one was administered during the last regularly-scheduled lecture in December, and part two was administered during the last scheduled laboratory session two days later. To guard against biased administrations, tests in both sections were administered by a laboratory assistant unfamiliar with FEHR or the evaluation project. To minimize the effects of scorer bias, the following procedure was used. Tests were numbered sequentially, then randomly shuffled by a secretary. Students were instructed to record the number of their test and not to write their names or other identification on the test paper. Names were assigned after scoring was completed.

Subscales. No a priori subscales were defined for this test. However, since FEHR experience seemed likely to affect some items more than others, each item score was recorded separately. Combination strategies are discussed in the analysis section. In subsequent discussions, the symbol FT is used to designate the total score on the final test.

Reliability. The tests were scored according to a detailed guide which allocated points for the presence of specific response characteristics. Using this guide, each test was scored independently by both the instructor and his teaching assistant. The correlation between the total test scores obtained under the two gradings was .87. For each item, the score assigned was the average of the two gradings rounded to the nearest whole number.

FEHR-PRACTICUM Product Rating Sheet

Description. The FEHR-PRACTICUM Product Rating Sheet (designated PRS for short) was developed by the project director to assess the proposals and reports produced in the practicum. It was intended to provide an objective measure of the quality of proposals and final reports for a broad range empirical investigations. Despite the title, the instrument is really more a checklist than a rating. The strategy adopted, wherever possible, was to identify elements or characteristics of proposals and/or reports that were both unequivocally identifiable and generally desirable. In developing the instrument we have been heavily dependent on material developed by Resta and Baker (1972) and Bruce W. Tuckman (1972), particularly with respect to organization and general content. However, the development of specific criteria required a number of rather arbitrary decisions for which the author takes sole responsibility.

A copy of the rating sheet appears on page 126. The rater's task is to enter in each of the eighty four blanks on the sheet a number from zero to the maximum value indicated (in parentheses) before each blank. The maximum value reflects the arbitrary weight assigned to that element in the overall assessment. The complete criteria for assigning numbers to each item are given in the guidelines which appears in appendix 4B. A few examples will be presented here to provide concrete illustrations of the scales involved.

FEHR-PRACTICUM RATING SHEET

for Proposals and Final Reports

Team ______ Members ______________

Date ______ Product ______________

A.Preliminary Materials		
1.Title page	3	☐
(a)precise prob. ident.	(2)	___
(b)suff. concise for index	(1)	___
(c)too long or wordy	(-1)	___
(d)incomplete author/info.	(-1)	___
2.Tables: Contents, figures, etc.	2	☐
3.Abstract	10	☐
(a)study purpose outlined	(2)	___
(b)target population identified	(1)	___
(c)major dependent variables	(1)	___
(d)design outlined	(2)	___
(e)analytic procedures outlined	(2)	___
(f)key comparisons outlined	(2)	___
B.Body of the Proposal		
1.Introduction	20	☐
(a)statement of the problem	(4)	___
(b)context or background	(4)	___
(c)purpose of proposed study	(4)	___
(d)importance of study	(2)	___
(e)scope and delimitation	(2)	___
(f)assumptions, limitations	(4)	___
(g)lacks logical relations	(-8)	___
2.Review of Related Literature	20	☐
(a)relation articles & study	(4)	___
(b)article methods evaluated	(4)	___
(c)articles representative	(4)	___
(d)logical grouping of studies	(4)	___
(e)results summar. & synth.	(4)	___
(f)critical studies are missing	(-8)	___
3.Conceptual framework(rationale)	24	☐
(a)set of principles or laws	(4)	___
(b)prin. tied to theory,research	(4)	___
(c)prin. form coherent unit	(4)	___
(d)prin. & modifying criteria	(2)	___
(e)research hypotheses stated	(4)	___
(f)definition of terms	(2)	___
(g)success crit.(objectives)	(4)	___
(h)lacks logical relations	(-10)	___
4.Method	40	☐
(a)subjects are described	(2)	___
-sampling described	(2)	___
-sampling representative	(4)	___
(b)design-described	(4)	___
-rationale	(2)	___
-variables not operat.	(-4)	___
-crit. compar. groups	(2)	___
-valid comparisons	(2)	___
-inval. not controlled	(-4)	___
(c)inst.-desc. all tests	(2)	___
-assess rel. & val.	(4)	___
-unsuitable, incomplete	(-4)	___

(d)data-source,who admin.,how	(2)	___
-data matrix defined	(2)	___
(e)analysis-rationale given	(4)	___
-covers hypotheses	(4)	___
-efficient	(4)	___
-inappropriate for purpose	(-10)	___
* 5.Budget	10	☐
(a)source of each item clear	(2)	___
(b)standard items present	(2)	___
(c)problems anticipated	(2)	___
(d)expense resource balance	(2)	___
(e)cost effectiveness assessed	(2)	___
* 6.Logistics	10	☐
(a)schedule of activities	(2)	___
(b)work distributed prorate	(2)	___
(c)sufficient personnel	(2)	___
(d)bottlenecks anticipated	(2)	___
(e)sequence logical & efficient	(2)	___
* 7.Personnel	10	☐
(a)major personnel named	(3)	___
(b)personnel respons. defined	(4)	___
(c)evidence of competency(vita.)	(3)	___
(d)personnel inadequate	(-10)	___
**8.Results(Statistical Concl.)	30	☐
(a)result for each hyp.	(4)	___
(b)explicit stat. concl.	(2)	___
(c)neat concise displays	(2)	___
(d)logical organization	(6)	___
(e)explan. graphs, diag.	(4)	___
(f)overall summary, synthesis	(12)	___
(g)procedural errors	(-10)	___
**9.Educational Conc. & Implic.	24	☐
(a)educ. meaning results given	(4)	___
(b)obj. not subj. presentation	(4)	___
(c)pattern of results interp.	(4)	___
(d)cost effectiveness assessed	(8)	___
(e)validity of concl. (target)	(4)	___
(f)misinterpretations	(-10)	___
10.Gen. Eval. of report/proposal	20	☐
(a)physically neat and orderly	(2)	___
(b)style acceptable(AERA,etc.)	(5)	___
(c)appropriate citations given	(3)	___
(d)organization clear,readable	(5)	___
(e)study is replicable	(5)	___
C.Supplementary Materials (bonus)	15	☐
1.Bibliography	(5)	___
2.Appended explanations of data	(10)	___

MAX. POSSIBLE/TOTAL ________/☐

RATING ______________

* Applicable to Proposal only

** Applicable to Final Report only

Most of the assignment rules are primarily quantitative. For example, the rule for section B.1(a) statement of the problem directs the scorer to:

Give:

4 points if there is an explicit statement of the "basic" or "root" problem. To rate full credit, the statement should identify, at least in general terms, each of the following:

(i) the system being studied.
(ii) what is presently happening in the system.
(iii) what should be happening in the system.
(iv) the reason for believing that it should happen.

Assign one point for each of the above elements present.

However, it was also recognized that the style and organization can reduce the communicative power of a presentation which contains all the elements of information to be communicated. Consequently, one or more elements in each section provide for subjective judgments of the cumulative negative effects of such flaws. For example, the assignment rule for section B.1(g) lacks logical relations directs the scorer as follows:

Give penalties of:

0 points if the material presented is smoothly connected and many of the above characteristics are present and individually meaningful, but there are inconsistencies, contradictions or ambiguities among characteristics.

-2 points if it would be necessary for the average member of the intended audience to read the section several times to determine what the study is about. (Do not impose this penalty if the re-reading is necessary because the reader does not have the background knowledge common to the writer's intended audience!)

-4 points if even after successive readings the average member would be uncertain about the study's purposes.

-8 points if after successive readings the average reader in the intended audience would have no idea what the study is about.

There was also a need to provide judgments on the organization, style, and readability of the proposal or report as an entity. A separate section (B.10) is devoted to that purpose. Again, the rationale was to upgrade objectivity by separating purely judgmental ratings from the "checklist" ratings wherever possible.

The question of element relevance proved difficult to handle. Obviously, the elements identified by certain items (e.g., B.1(f) assumptions and limitations) were irrelevant and unnecessary in some studies. Yet, to allow each rater subjectively to determine whether each element was relevant to a particular study would certainly decrease the objectivity of the scale. Two actions were taken to minimize the effects of item relevancy on overall quality scores and on the relative weightings placed on various sections of the document. First, wherever it was logical to expect substantial numbers of studies for which an item was irrelevant, the scoring instructions began with an award of full points and subtracted points for relevant data which was missing rather than adding for adata which was present. Second, in a case where a particular element was clearly beyond the scope of the writer's responsibility (e.g.,in a FEHR problem which specifically excluded a review of the literature), the expected score was amended to zero. Thus, for any section, one could calculate both an absolute score based on the information which was present and a relative score based on the proportion of the assigned tasks which were present.

Role. The product rating scale was considered to be absolutely vital to the evaluation because it was based on assessment of an element common to all research/evaluation activities: the research product. It had the additional advantage that proposals and reports formed permanent records, thus permitting scoring procedures to be reviewed and verified. There was, however, an important disadvantage to the strategy. In addition to the amount of time required to develop the instrument (about three months), a conscientious scoring of one proposal or final report took at least two hours: approximately thirty-eight man/days of labor on just this phase of the evaluation. On balance, the increased information was considered adequate justification for the time investment.

The instrument was developed during the winter of 1973, and formatively evaluated during that summer and fall using all the products from the 1972-73 year. The scoring criteria gradually evolved over that period, reaching its present form by December. The instrument was used to score all proposals and reports collected in the FEHR-PRACTICUM field trials. Because of the changes during development, it was necessary to rescore all the 1972-73 materials. To obtain a comparative base from which to judge adequacy, the product rating sheet was also used to score nine randomly-selected dissertation proposals (class 20).

Subscales. A total of eighty four separate item scores plus fourteen subtotal scores (as indicated by the rectangular boxes) were available from the rating scale. However, for purposes of this evaluation, some of the subtotals were grouped into larger summary scales. The major summary scales to be used here are:

1. Introduction and Problem Definition Scale (IP): the sum of items B.1(a) to B.1(g).
2. Review Scale (RV): the sum of items B.2(a) to B.2(f).
3. The Conceptual Framework Scale (CF): the sum of items B.3(a) to B.3(h).
4. The Method Scale (M): the sum of item B.4(a) to B.4(e).
5. The Logistics Scale (LG): the sum of items B.5(a) to B.7(d).

6. The Result/Conclusion Scale (RC): the sum of items B.8(a) to B.9(f).

7. The General Evaluation Scale (GE): the sum of items B.10(a) to B.10(e).

8. The Composite Scale of Commonly Assigned Proposal Elements (CP): the sum of the IP, CF, M, and GE scales. Since these elements were common to both proposals and reports, it was useful for comparing groups which had completed one or the other but not both a proposal and a final report. The R scale, which one would normally want to include in this composite scale, was excluded here because of the tremendous variety among evaluation sites in the resources and expectations for this element.

9. The Composite Scale of Commonly Assigned Final Report Elements (FC): the sum of the IP, CF, M, RC, and GE scales.

10. A PP Scale representing the proportion of the assigned (or expected) proposal tasks credited was calculated by dividing the total of all proposal tasks for each subject by the total possible score for his class if all assigned tasks had been satisfactorally completed.

11. A PF Scale representing the proportion of the assigned (or expected) final report tasks credited was calculated by dividing the total of all final report tasks for each subject by the total possible score for his class if all assigned tasks had been satisfactorally completed.

Rater Reliability. The ideal procedure for estimating rater reliability would have been to insert exact replicates of previously-scored documents at random intervals throughout the data. This plan was rejected because of the amount of time involved. However, a rough estimate of the minimum value of the random replicates reliability could be obtained by what was labelled the identical-elements correlation.

Each of the seventy-odd subjects in class 14 (C655, fall 1973) were required to write a formal proposal for evaluating the effectiveness of the various treatments which involved one assigned dependent variable, one assigned moderator (independent) variable, and one "personal-interest" variable (chosen by the subject). The students worked together in teams of three to develop their proposals, but the assignments were arranged so that the members of every team had two variables in common. Because of this feature, the proposals developed by members of the same team had many identical elements which were developed cooperatively.

The procedure used was to (temporarily) mask the identification on each proposal, mix proposals from all classes thoroughly, then complete the rating sheets for the entire set. When this task was complete, the products were identified and matched by team membership into all possible pairs. For each pair the scores on all identical elements were computed and the correlation between paired scores computed over the entire class. Because of the redundancy in the within-team pairing procedure, the degrees of freedom for the correlation were defined by (total degrees of freedom within teams = 2) rather than (number of pairs = 2).

The correlation obtained by this process was .6759. Since the "identical" elements were seldom as much as half the paper and included errors attributable to differences in format and style as well, this was considered a respectable level of rater reliability.

Self Assessment of Research and Evaluation Skills (SARES)

Description. The SARES instrument contained forty items describing tasks such as: "compute and interpret a one-way analysis of variance" and "distinguish among main effects, interactions, simple main effects, and confounded effects." The subject was asked to provide three ratings for each task: (1) their competence to perform the task, (2) their interest in that sort of task, and (3) the importance or relevance of the task for their planned career. A five-point scale was used for each rating, with one representing complete absence of the characteristic being assessed, and five representing a superior level. A copy of the instrument appears in appendix 4C.

Role. The instrument was designed to be used with all field-study subjects to measure the effects of various types of experience on subjects' perceptions of their competence, interest and relevance with respect to various research tasks. The instrument was administered to the experimental evaluation classes (1-3) at the end of the semester (December 1972), and to the same people again at the end of the second semester in May 1973. For these groups the questionnaire seemed adequate. However, when it was administered to some sample control subjects, they found the language so technical and unfamiliar that they were unable to respond. Consequently, another instrument -- the FEHR questionnaire -- was developed to permit control group comparisons.

Subscales. Within the competency interest and relevance dimensions, the forty item responses were grouped into eight content areas: (1) elementary statistics, (2) senior statistics, (3) sampling, (4) scaling, (5) measurement, (6) design, (7) goal explication, and (8) completing a dissertation. In addition, mean competency (MC), interest (MI), and importance or relevance (MR) scales were computed by averaging over the eight content areas. These three scales were used for all between-group comparisons.

Reliability. In the summer of 1972, the instrument was administered to a small class (N = 22) on two difference occasions. Since there was only a one-week interval between administrations, it was assumed that no actual changes in attitude had occurred and errors could be attributed to unreliability in the test. The obtained correlations among the mean competency, interest and importance scores at first and second administrations appear in figure 4.3.

Goal Assessment Questionnaire (GAQ)

Description. The GAQ was used as a course-evaluation device by all the FEHR classes conducted at the University of Michigan. It differs from the usual course evaluation questionnaire in that each student response is explicitly related to an instructional objective and a corresponding criterion of success. In this study, all responses were related to the following nine goals and their associated criteria.

	Competency Time 1	Interest Time 1	Importance Time 1	Competency Time 2	Interest Time 2
Competency Time 1	——				
Interest Time 1	.4795	——			
Importance Time 1	.4374	.7124	——		
Competency Time 2	.7375	.4713	.4513	——	
Interest Time 2	.4825	.7198	.6455	.4998	——
Importance Time 2	.4692	.6537	.7592	.5294	.7077

Critical r @ .95 level = .4227; Critical r @ .99 level = .5368

Figure 4.3. Matrix of Correlations Among Mean Competency, Interest, and Importance Ratings Collected One Week Apart.

GOAL 1: To increase interest in research methodology and confidence in dealing with statistics and statistics-related courses by using FEHR problems.

Criterion. This goal has been moderately achieved if you found this course to be more interesting and/or less anxiety-arousing than you had expected. It has been completely achieved if you are now sufficiently interested and confident to enjoy learning about new research methods.

GOAL 2: To develop an increased appreciation for the innate complexity of evaluating program effectiveness.

Criterion. This goal has been achieved if you now consider more dimensions in the evaluation task than previously (e.g., use more variables to "measure" the effects).

GOAL 3: To develop the tolerance for ambiguity and patience necessary to deal with a complex problem.

Criterion. You have completely achieved this goal if, when given a problem for which no final solution is evident, you proceed with any method which will reduce the uncertainty -- even by a small amount -- confident that a step-by-step approach will eventually lead to a solution.

GOAL 4: To integrate and interrelate your existing knowledge and skills in measurement, research design, statistics, psychological theory, and educational practice.

Criterion. You have achieved this goal if you now feel that your knowledge and skill in two or more of these areas somehow "make more sense" or "fit together better."

GOAL 5: To attain sufficient skill in data analysis to compute up to a t test by hand (i.e., using only a calculator), and to interpret the results.

Criterion. You have achieved this goal if you have correctly followed the cookbook formulas for this task, shown where the numbers come from and interpreted the results of a computer analysis of the data.

GOAL 6: To develop the ability to use an appropriate computer program to do more complex analyses such as a one-way ANOVA with subsequent comparisons using the combination command and to interpret the results.

Criterion. You have achieved this objective if you have successfully completed both of the above designs on sample data, and were able to interpret the results.

GOAL 7: To develop the ability to identify the common threats to internal and external validity when both are present in a study, and to suggest research techniques to control these threats.

Criterion. If you are confident that you can recognize examples of any of the common sources of invalidity (in an open book situation), and can suggest some method of controlling each threat, you have completely achieved this objective.

GOAL 8: To develop the ability to state a (given) research/evaluation problem in terms of relationships among variables.

Criterion. If you have participated in a successful FEHR proposal and are confident that you can state a new problem in operational terms, you have achieved this objective.

GOAL 9: To develop the ability to write research/evaluation proposals and final reports in an acceptable style.

Criterion. If you have participated in writing an acceptable proposal and report, and are confident that, with minor help, you could complete the task by yourself, you have completely accomplished this objective.

For each goal listed above, the subjects were asked to respond to the following six questions by choosing one of the five options provided.

1. In terms of my professional development, the attainment of this instructional goal is:
   (1) Likely to detract from my professional performance.
   (2) Unrelated to my professional performance.
   (3) Necessary for masterful performance but not for adequate performance.
   (4) Necessary for adequate professional performance.
   (5) A prerequisite which must be mastered before adequate performance can be developed.
2. During the course, the instructor's communication of this goal to our class was:
   (1) Essential but not attempted.
   (2) Useful (rather than essential) but not attempted.

    (3) Not attempted and not needed.
    (4) Attempted but needed further clarification.
    (5) Clear and adequate.

3. Within the time constraints of the course, the task of achieving this goal was (is) for me:
    (1) Very easy. Accomplished with very little effort. Accomplished before the course began.
    (2) Moderately difficult. Accomplished with moderate effort.
    (3) Difficult. Accomplished with considerable effort.
    (4) Extremely difficult. Accomplished only with great effort.
    (5) Impossible for me to accomplish in the time available.

4. Regardless of the difficulty or ease indicated, my achievement of this goal, in terms of the criteria suggested by the instructor, is:
    (1) Well below criterion performance.
    (2) Somewhat below criterion performance.
    (3) Close to criterion performance, but some question remains.
    (4) Clearly adequate, at or somewhat above criterion performance.
    (5) Well above criterion performance.

5. During this course, assignments and/or laboratory exercises which provided an opportunity to achieve this goal were:
    (1) Necessary, but not provided.
    (2) Not provided, but unnecessary.
    (3) Present, but more were needed.
    (4) Present in adequate quantities.
    (5) Present in quantities greater than warranted.

6. The emphasis placed on this goal, relative to other course goals, should be:
    (1) Greatly decreased.
    (2) Decreased somewhat.
    (3) Left as it is.
    (4) Increased somewhat.
    (5) Greatly increased.

Role. The GAQ was administered to all FEHR classes at the University of Michigan to provide a direct measure of student perceptions of the degree to which they had achieved the instructional goals listed previously. It was not used in the off-campus trials because it was not possible to obtain a priori concensus on the goal statements.

Subscales. For purposes of this study the questionnaire results were reduced to two scales per goal. The first was the goal importance as measured by the mean of questions 1 & 2 and the second was goal attainment as measured by the mean of questions 3 & 4.

Reliability. Neither internal consistency nor test-retest reliability estimates were considered adequate for this study. The former was unsatisfactory because the items were designed to measure attributes which were conceptually independent. The latter was unsatisfactory since we were predicting treatment effects which were interactive with organismic variables such as ambition or activity level. However, reliability is of value primarily because it is a prerequisite to validity. Since the study results constitute evidence for the construct validity of the GAQ subscales, no estimate of reliability was considered necessary.

ORS Questionnaire

Description. The ORS questionnaire consisted of six semantic differential ratings on each of eleven different elements of a research enterprise. The subject's attitude towards an element was defined by placing an X in one of a series of blanks separating six pairs of polar adjectives. The first item is listed with the X's placed to indicate a great need for research skills combined with fear of mathematics.

(1) Statistics is:

(a)	Intimidating	X :__:__:__:__:__	Inspiring self-confidence
(b)	Irrelevant to my future work	__:__:__:__:__: X	Necessary for my future work
(c)	Wearisome	__:__:__: X:__:__	Interesting
(d)	Conceptually difficult	X :__:__:__:__:__	Conceptually simple
(e)	Complex in practice	X :__:__:__:__:__	Simple in practice
(f)	Unrewarding	__:__:__: X:__:__	Satisfying

The remaining elements to be rated on these six dimensions are listed below, with the bipolar adjectives omitted to save space.

(2) Computers
(3) The research process
(4) Research design
(5) Defining successful completion of an educational objective
(6) Proposal writing
(7) Identifying the basic need a proposed program is trying to meet
(8) Basing decisions on research
(9) Budgeting time, money and other resources
(10) Practicum experience in research
(11) Team work in research

Role. As mentioned previously, the rating scale developed for the experimental evaluation proved too technical for research novices. This instrument was intended for use with both experienced researchers and the research novices from the control groups. It was administered to all classes except the first three. (Note, however, that most of the students from these classes did respond to the questionnaire during their second semester of research training; that is, in classes four to six.)

Subscales. The sixty six ratings resulting from the questionnaire were reduced to ten subscales, as follows. First, the six ratings on each element was to an interest dimension computed by averaging of ratings (a), (b), (c) and (f), and a difficulty dimension computed by averaging ratings (d) and (e). The data was further reduced by combining the original eleven elements into five categories, with each category having an interest scale and a difficulty scale. The five categories and the corresponding scales were:

1. (1) The classical research elements (items 1, 2, 3 and 4) which produced the ICR and DCR scales.
2. (2) The program evaluation elements (items 5, 7 and 8) which produced the IPE and DPE scales.
3. (3) The proposal writing elements (items 6 and 9) which produced the IPW and DPW scales.
4. (4) The research practicum elements (items 10 and 11) which produced the IRP and DRP scales.
5. (5) The grand mean of all items which produced the MI and MD scales.

Reliability. The questionnaire was administered pre and post to a number of classes. Unfortunately, the confidentiality requirements enforced on some sites made it impossible to pair the pre and post scores at many sites. The correlations for the 47 people for whom pre-post pairings could be firmly established appear in figure 4.4.

Dimension	Scale				
	Classical Research	Educ. Eval.	Proposal Writing	Research Practicum	Grand Mean
Interest	.4289	.3798	.2094	.0070	.2853
Difficulty	.4531	.1766	.1540	.3207	.2609

Figure 4.4. Pre-post Correlations for ORS Subscales.

The wide disparity in correlations does not mean that "interest in the research practicum" is less reliable than "interest in classical research." In fact, it is probable that the difference in correlations occurs because the FEHR experience caused a moderate positive increase in the "classical" interest for most people, while the "practicum" interest was radically improved for some people, mildly improved for the majority, and radically decreased for a small minority. Non additive treatment effects result in reduced correlations. It seems reasonable to assume that the highest obtained correlation (.45) represents the minimum bound of reliability. However, the assumption is not critical: the legitimacy of the scales derive from the construct-validity evidence which is implicit in the results obtained.

Summary of Instrumentation

An overview of the instruments used with each class is provided by figure 4.5. An X indicates that the instrument named at the top of the column was administered to the class listed in the left hand margin.

SECTION II. EMPIRICAL EVIDENCE

The previous section presented an overview of the data sources to be used in the summative evaluation of FEHR-PRACTICUM. In this section we shall present the empirical data on which the evaluation is based. It is convenient to discuss the experimental evaluation and the field evaluation separately under the labels study one and study two. Within each of the studies the material is organized as follows:

(1) A brief introduction describing the role of the study in the total evaluation followed by statements of:
   (a) The specific purposes of the study.
   (b) The rationale upon which the study is based.

(2) A description of the experimental method under the following headings:
   (a) Subjects and sampling plan
   (b) The educational treatments
   (c) The design, including:
      - a diagrammatic summary

CLASS	N	FIRST EXAM. (E[illegible])	FINAL EXAM. (FT)	SARES QUEST.	PROJECT RATINGS		GOAL RATINGS (GAQ)	ORS QUEST. (ORSQ)
					PROPOSALS	FINAL REPORTS		
Experimental Evaluation								
1. C655 (UM) Sec. 1, F 72	25	x	x	x	x	x		
2. C65[illegible]M) Sec. 2, F 72	22	x	x	x	x	x		
3. C6[illegible] Sec. 2 & K680, F 72	6	x	x	x	x	x		
4. C65[illegible], W 73: (a) S's from (1)	15			x	x	x	x	x
(b) New S's	18			x	x	x	x	x
5. C656, S's from (2), W 73	21			x	x	x	x	x
6. C656, S's from (3), W 73	6			x	x	x	x	x
Field Trials								
7. Mich. State U., W 73	19				x	x		x
8. Ohio State U., W 73	13				x	x		
9. West. Mich. U., W 73	5				x	x		
10. Indiana U., W 73	8				x	x	x	x
11. C699 (UM), F 72	6				x	x	x	x
12. C699 Flint, W 73	7				x	x	x	x
13. C699 (UM), W 73	6				x	x	x	x
14. C655 (UM) Sec. 1, F 73	46				x		x	x
15. C655 (UM) Sec. 2, F 73	30				x		x	x
16. C656 (UM) Sec. 1, W 74	32						x	x
17. C656 (UM) Sec. 2, (a) W 74	19						x	x
(b) New S's	4						x	x
FIELD TEST TOTAL N	308							
Control Groups: Field Trials								
18. A705 (UM), W 73	30							x
19. C607/C654, W 73	11							x
20. Ph.D. Proposals (sampled randomly from Ed. Psych. files).	9				x			
GRAND TOTAL N	358							

Figure 4.5. Summary of Instrumentation. An X indicates that the instrument named in the column heading was administered to the class identified in the row heading.

        - dependent variables
        - moderator and control variables
    (d) The research hypotheses
    (e) The analytic plan
(3) The results of analysis
(4) A summary of the findings

STUDY ONE: EXPERIMENTAL EVALUATION

It is obvious from the summary of participating classes displayed in figure 4.1 that the experimental evaluation was actually a subset of the field evaluation rather than an independent study. It consists of that portion of the evaluation for which it was possible to do some random assignment and to exercise a modicum of control over treatments. More specifically, we shall be concerned with the information obtained from the first six classes.

Purpose. The specific purpose of this study was to assess the effectiveness of FEHR-PRACTICUM as a laboratory experience to accompany graduate education courses in research design and data analysis. In particular, we wished to compare the achievement and research attitudes of students given FEHR projects and students given the traditional skill practice using encapsulated data from prior studies (i.e., printed problems).

Rationale. Students in graduate education, regardless of their specialty, are usually required to develop a "research competency." In practice this has traditionally meant that they were required to complete one or more courses in the area of measurement, research design, and data analysis. It is frequently the case that a large proportion of the students entering these courses are there only because of the requirement. They typically have had little mathematical training beyond high school (often many years ago) and are fearful of the statistical content. In addition, a substantial number of students consider both statistical theory and practical laboratory experiences based on neat "canned" experiments to be largely unrelated to the work for which they are being trained. Since a

FEHR-PRACTICUM project requires the participants to apply research methods to a practical on-going problem, it should help bridge the gap between theory and practice. In addition, the intense involvement which has been characteristic of FEHR participants in the past ought to ameliorate the effects of fear of statistics. If this in fact happens, one would expect that the earlier the exposure to FEHR, the better, and that increasing the amount of exposure (either through more complex problems, or a greater number of problems, or both) will result in increased achievement and an improved attitude toward research.

Method

Subjects. The subjects for the study consisted of all students in Education C655 at the University of Michigan during the fall semester of 1972. Education C655 is the first in a two semester sequence of research design and data analysis courses which are required by most of the graduate programs in education as evidence of research competency. It is typical of similar courses in other colleges of education in that only a small minority of these students have had previous research experience or a college course in mathematics.

The students had registered for C655 with the understanding that the class would be split into two approximately equal sections which would share a Monday lecture session, but have separate laboratory sections to be scheduled on Wednesdays and Thursdays. The two laboratory sections comprised the control and experimental treatments. Unfortunately, it was not possible to assign students to laboratory sections entirely at random. Twelve students had schedule conflicts which required a Wednesday laboratory and eight required a Thursday assignment. In addition, there were six students who were specializing in research methods and were therefore simultaneously enrolled in a C699 course which involved an intensive FEHR project. The latter were considered a unique group. The remainder of the students were randomly assigned to sections so as to obtain equal numbers of students. On the flip of a coin, the Wednesday section was assigned to the experimental FEHR condition, and the

Thursday section to the control condition. Luckily, the six students simultaneously enrolled in C699 were all able to meet either day. Rather than introduce a condition which mixed FEHR students with students experiencing the traditional laboratory practice, the six were assigned to the experimental condition. Thus, two levels of involvement within the experimental section were created. The regular experimental laboratory was designated a restricted FEHR involvement, and the experimental laboratory supplemented with a C699 project was designated an extensive FEHR involvement.

Although the Wednesday and Thursday laboratory sections were initially of equal size, there were four late registrants and one drop. The final distribution was 28 subjects in the Wednesday experimental section (22 restricted and 6 extensive), and 25 subjects in the Thursday control section.

Since most of the students who enrolled in Education C655 also enrolled in the sequential C656 course the following semester, the experiment was planned to continue over two academic terms. When winter enrollments were stabilized, the student body of Education C656 consisted of 15 students from the fall control group, 21 students from the experimental restricted involvement level, 6 from the experimental extensive involvement level, and 16 new students who had taken their first-semester course elsewhere. The distribution of subjects is summarized in figure 4.6.

Group	Term 1 Education C655	Term 2 Education C656
Control	n = 25	n = 15
Experimental Restricted	n = 22	n = 21
Experimental Extensive	n = 6	n = 6
New Students for C656		n = 15

Figure 4.6. Distribution of Subjects in the Experimental Evaluation.

Treatments. Before proceeding with a technical discussion of the design, it is useful to have a better understanding of the substantive content and instructional strategies involved in the various course sections which constitute the educational treatments in the experiment. A brief description of each treatment is provided here, with supporting details provided in the referenced appendices. The discussion is organized chronologically into fall term and winter term treatments.

Fall Term: Education C655. The content for which all students were responsible was presented in a two-hour lecture session on Monday exclusively. This session was attended by students from both sections. The content of the course was classical research design and data analysis with little emphasis on problem definition or decision oriented research. Topics included basic research design, descriptive statistics and inferential statistics up to a one-way analysis of variance. The treatments to be compared were the different laboratory sessions. These are described below under three headings: control, experimental, and experimental double exposure.

1. Control. Students in the control section were given laboratory problems which required the practical application of the principals studied in class to new data sets in laboratory handouts. To encourage generalization to practical situations, many of these were presented as synopsized research projects. However, no attempt was made to provide continuity from project to project. The entire set of laboratory exercises appears in appendix 4D. The instructor and his teaching assistant circulated about the laboratory helping students complete their assigned exercises. On October 15, when the experimental section was formed into teams, the control group was also formed into three-man groups. These groups worked on the laboratory problems collectively, but each person was expected to complete all exercises and to be able to explain what was done.

2. Experimental Restricted. Students in both the experimental sections were given the same laboratory problems as the control group until October 15. On that day they were given synopses of the eight FEHR-PRACTICUM problems and asked to form research teams of two or three members each. Each team was to complete a FEHR-PRACTICUM project for the problem of their choice before December 11. For the remainder of the term, they were told, all laboratory sessions were to be spent on their project. The groups worked collectively on their project, but each individual was expected to produce his own report and to be able to explain the rationale for each step in their experiment and to discuss the educational meaning of each finding. The specific tasks assigned were defined by the checklist of tasks, as discussed previously.

3. Experimental Extensive. As a supplement to the regular course project described above, every student had the option of doing a complete FEHR-PRACTICUM project for extra credit via a companion course Education C699, which could be elected simultaneously. This course did not involve any additional instruction or a scheduled class session: it was a vehicle for awarding credit for intensive team research efforts. Students who enrolled in this course were expected to spend several additional hours each week (over and above their C655 commitment) in "solving" their FEHR problem. The C699 project was much more comprehensive than the one assigned in C655. It involved submitting formal proposal (including budget) and negotiating funding as well as completing an evaluation project and writing a formal report. However, students who were simultaneously enrolled in the two courses completed a single report for the two classes. The extra time did not provide an unfair advantage over other C655 students since the course grading procedures were kept independent of the project products per se, as demonstrated by the final examination.

Winter Term: Education C656. The topics covered in the second course included intermediate research design (blocking, balancing, etc.), factorial of variance, one-way and factorial analysis of covariance, the general multiple regression, non-parametric statistics, and an introduction to multivariate techniques. A complete syllabus is available on request. The course organization was similar to the first semester, with two laboratory sections (Wednesday and Thursday) and a common lecture session (Monday). Both second-term laboratory sections featured an intensive FEHR project: no control type of laboratory was offered. However, students again had the option of enrolling in Education C699 for additional FEHR experience. Thus there were again two levels of involvement. These were designated intensive and intensive/extensive respectively.

4. Experimental Intensive. The regular C656 conducted in essentially the same fashion as the C655 experimental laboratory, but with three modifications: (1) the project lasted the entire semester, (2) a great deal more research sophistication was expected, and (3) the proposals and final reports were included in the course grading system. The number of tasks assigned was the same as for the fall C699 project, but the students increased knowledge of sources of invalidity and the possibilities for statistical control produced an approach that was intensely concentrated: hence the distinguishing label.

5. Experimental Intensive/Extensive. As mentioned, the experimental intensive/extensive group took a C699 course simultaneous with C656. However, unlike its fall counterpart, this course required each student to complete a comprehensive FEHR project on a different problem than the one used in C656. Thus, each member of the experimental intensive/extensive group produced two proposals and two final reports during the winter term. Both of these were intensive projects.

Sampling	Sept. 11	Oct. 16	Oct. 16	Dec. 11	Jan. 8	Apr. 30
PR	X_0	O_{11}	X_0	O_{21}	X_3	O_{31a}
					X_4	O_{33a}
PR	X_0	O_{12}	X_1	O_{22}	X_3	O_{32}
					X_4	O_{33b}
IG	X_0	O_{13}	X_2	O_{23}	X_3	O_{33c}
					X_4	O_{33d}
IG	New students beginning in C656				X_3	O_{31b}
					X_4	O_{33e}

Figure 4.7. Schematic Representation of the Experimental Design.

Design. The design of the study is represented schematically in figure 4.7 using a notation adapted from Campbell and Stanley (1963). The symbols in the left hand column define the sampling procedures described in the previous section. The letters PR and IG stand for partially randomized and intact groups respectively. The letter X represents a treatment which began on the date appearing above the column in which it appears. The subscript identifies both the particular treatment or combination of treatments administered and the degree of FEHR exposure (the higher the number the greater the exposure), as shown by the following key:

0 Fall control treatment. A laboratory problem in applied statistics was completed each week. A different problem was used each week.

1 Fall experimental restricted. The assigned FEHR project required only a final report. No proposal was required, and costs were ignored throughout the project.

2 Fall experimental extensive. A C699 class was taken simultaneously with the regular C655: An extensive FEHR project

lasting all semester was required. It featured both a proposal (complete with literature review and budget) and a final report. This project also counted as the C655 laboratory project.

3 Winter experimental treatment. An extensive FEHR project lasting all semester, and requiring both a proposal and a final report. The proposal required a budget, but only the information bank literature was reviewed.

4 Winter C699 class was taken in addition to the regular C656 class. Students taking this course completed a second complete project in addition to the regular C656 project.

The letter O represents a set of observations taken during the week which begins on the date above the column. The first subscript identifies the time at which the measurement is taken and the second identifies the group which was observed. The subscripts a, b, c, d, and e in the last column are used to indicate subgroups. These five subgroups were pooled, after appropriate testing for similarity, to form the larger treatment groups. The procedure for doing this is described in the analysis section.

In the remainder of this section a dot (.) is used for a subscript to indicate that data has been pooled over the elements identified by the subscript concerned. Thus, $O_{1.}$ refers to the set of measures taken from all groups at time 1, and $O_{.2}$ refers to all the measures taken on group 2 during the whole experiment (i.e., at all three times). This notation is particularly convenient for the operational specification of critical comparisons. The variable scores contained in each observation set, and the instruments from which they derive appear in figure 4.8.

Hypotheses. The hypotheses to be tested in this study derive directly from the first five general objectives of the summative evaluation. If the FEHR system is meeting its objectives, then an increase in exposure (either the complexity (intensity) of a project or in the number of projects completed) ought to produce a monotonic increase in the variables which operationally define the objective.

SET	INSTRUMENTS ADMINISTERED	SUB-SCALE	SYMBOL	FUNCTION
O_1.	First Examination: C655	Total	E1	Covariate
O_2.	Final Examination: C655	Total	FT	Dep. Var.
	Self Assessment of Res. and Eval. Skills (SARES)	Mean overall for		
		1. Competency	MC	Dep. Var.
		2. Interest	MI	Dep. Var.
		3. Relevance	MR	Dep. Var.
	FEHR-PRACTICUM Product Rating Sheet (PRS)	1. Intro. & Prob. Def.	IP	Dep. Var.
		2. Review of Lit.	RL	
		3. Conceptual	CF	
		4. Method	M	
		5. Logistics	LG	
		6. Results	RC	
		7. Evaluation	GE	
		8. Proposal	P	
		9. Proportion Prop.	PP	
		10. Final Rpt.	F	
		11. Proportion Final Rpt.	PF	
		12. Proposal Composite	PC	
		13. Final Composite	FC	
O_3.	Self Assessment of Res. and Eval. Skills (SARES)	Mean gain from O_2.		
		1. Competency	GC	
		2. Interest	GI	
		3. Relevance	GR	
	FEHR-PRACTICUM Product Rating Sheet (PRS)	As above	As above	Dep. Var.

Figure 4.8. Variables Measured at Each Observation Time.

The degree of FEHR exposure involved in the various treatment combinations is expressed symbolically below. For convenience, we have ignored the first five-week control treatment administered to everyone at the beginning of the experiment.

$$X_0 < X_1 < X_2 < X_3 < X_4$$

The corresponding relationship among observation sets is given by:

$$O_{21} < O_{22} < O_{23} < O_{31} < O_{32} < O_{33}$$

The specific hypotheses to be tested in this study can be generated by stating each of the first five objectives as a major substantive hypothesis, specifying the scales which operationally define the dependent variables of interest, and then stating the specific hypotheses as expected relationships among the means of observation sets for each scale. The scale symbol is subscripted to identify the observation set as described in the design, and a bar over a symbol signifies the mean score of the observation set concerned.

From objective one it was hypothesized that an increased exposure to FEHR would produce a monotonic increase in achievement and perceived achievement. Achievement at observation time two was operationally defined as the total score on the final examination (FT). Perceived achievement at both times two and three was defined as the mean competency scale (MC) from the SARES instrument.

The immediate effects of the experimental treatment could be assessed by comparisons within time two. If the accured advantage persisted into the next term, one would expect a similar trend among the difference between time three and time two scores (DC). (The MC scores from time three reflect both time two and time three differences: therefore, the difference score was used.) Finally, if the trend continued one would expect a monotonic increase in MC scores with increased experience over the whole experiment: i.e., over both time two and three. However, the design did not protect against the spurious effects of repeated testing: consequently, support of the first two hypotheses and non-support of the overall trend would not constitute a negative finding. On the other hand, a consistent monotonic increase would tend to be supportive of the underlying theory.

Summarized symbolically by scales, the four hypothesized relationships were:

(1) $\overline{FT}_{21} < \overline{FT}_{22} < \overline{FT}_{23}$

(2) $\overline{MC}_{21} < \overline{MC}_{22} < \overline{MC}_{23}$

(3) $\overline{DC}_{31} < \overline{DC}_{32} < \overline{DC}_{33}$

(4) $\overline{MC}_{21} < \overline{MC}_{22} < \overline{MC}_{23} < \overline{MC}_{31} < \overline{MC}_{32} < \overline{MC}_{33}$

From objective four it was hypothesized that at time two an increased exposure to FEHR would produce a monotonic increase of interest in research and research methods as measured by the MI scale (mean of all the interest ratings) on the SARES instrument. A similar increase in the time three minus time two difference scores (DI) was hypothesized. In addition, it was hypothesized that MI scores would show a monotonic increase with FEHR exposure over the entire experiment. However, MI scores, like MC scores are subject to the effects of testing. Again, non-support of the overall trend would not by itself constitute negative evidence. Summarized symbolically by variable, the hypothesized relations were:

(5) $\overline{MI}_{21} < \overline{MI}_{22} < \overline{MI}_{23}$

(6) $\overline{DI}_{31} < \overline{DI}_{32} < \overline{DI}_{33}$

(7) $\overline{MI}_{21} < \overline{MI}_{22} < \overline{MI}_{23} < \overline{MI}_{31} < \overline{MI}_{32} < \overline{MI}_{33}$

From objective five it was hypothesized that an increased exposure to FEHR would produce a monotonic increase in the perceived relevance of research as measured by the MR subscale (mean of all importance ratings) from the SARES instrument at time two and a similar monotonic increase in time three minus time two difference scores (DR). It was also hypothesized that there would be a monotonic increase in MR over both time periods. However, because of the possible effects of testing, non-support of the last hypothesis would not by itself constitute negative evidence. Summarized symbolically by variable, the hypothesized relations were:

(8) $\overline{MR}_{21} < \overline{MR}_{22} < \overline{MR}_{23}$

(9) $\overline{MR}_{31} < \overline{MR}_{32} < \overline{MR}_{33}$

(10) $\overline{MR}_{21} < \overline{MR}_{22} < \overline{MR}_{23} < \overline{MR}_{31} < \overline{MR}_{32} < \overline{MR}_{33}$

From objective two it was hypothesized that an increased exposure to FEHR would produce a monotonic increase in the quality of proposals and dissertation as measured by the CP and CR score patterns from the product rating sheet. The CP pattern for each individual consisted of the scores on each of the sections commonly assigned for proposals (IP, CF, M, GE), and the CR pattern consisted of the scores on all sections commonly assigned for a final report (IP, CF, M, RC, GE). Summarized symbolically by patterns, the two hypothesized multivariate relations were:

(11) $\overline{CP}_{22} < \overline{CP}_{23} < \overline{CP}_{31} < \overline{CP}_{32} < \overline{CP}_{33}$

(12) $\overline{CR}_{22} < \overline{CP}_{23} < \overline{CR}_{31} < \overline{CR}_{32} < \overline{CR}_{33}$

From objective three it was hypothesized that an increased exposure to FEHR would produce a monotonic increase in the quality of field study designs. Quality of design was operationally defined by the method (M) scale from the product rating sheet. Summarized symbolically, the hypothesized relation was:

(13) $\overline{M}_{22} < \overline{M}_{23} < \overline{M}_{31} < \overline{M}_{32} < \overline{M}_{33}$

Analytic Plan. The data analysis was conducted in two parts. First the data from time two was analyzed to test the direct effects of the experimental laboratory sessions, and then the data from both times two and three was analyzed to test for pervasive overall trends.

Analyses of Semester One Scores (O_2). Four scales were analyzed at time two: FT, MC, MI, and MR. The FT scores which represented achievement were analyzed separately from the three scores from the SARES instrument because of the different scales involved.

In both analyses the scores on the first examination (FE) were used as the covariate to correct for initial differences in statistical ability. The monotonicity hypothesis was tested in each case by two planned tests on trends: a test of linear trend and a test of non-linear (curvelinear) trend. All analyses were conducted by computer using the fully documented Michigan Interactive Data Analysis System developed and tested by the Statistical Research Laboratory at The University of Michigan. The steps followed in conducting each analysis are specified below:

1. FT Scores. An analysis of covariance of the FT scores stratified by exposure levels was conducted, followed by orthogonal contrasts to test for linear and curvelinear trends in means with increasing FEHR exposure.

2. SARES Scores. Only the three grand-mean scales (MC, MI, and MR) were analyzed. The first step was to compute a profile analysis to test whether the score profiles were parallel for the three exposure groups.

   Since neither a covariance analysis nor orthogonal comparisons were available in the profile analysis, the three variables were entered into separate analyses of covariance followed by orthogonal tests for linear and non-linear trends.

3. Overall Analysis. Obviously, multivariate procedures were most appropriate for analyzing both the SARES scales and the multiple dependent variables derived from the product rating sheet (PRS). However, multivariate procedures for analyzing repeated blocks of measures were not available. Consequently we were forced to choose one of two alternative procedures. We could either analyze via a series of univariate analyses of covariance, or we could ignore the repeated measures aspect (from time two to time three) and conduct a multivariate analyses. Either procedure could be followed by orthogonal tests for linear and non-linear trends. According to our plan, the former procedure was to be used if the covariate correction for time

two scores proved to be significant ($p > .05$), and the latter was to be used if it was not significant. Since the covariate did not significantly affect the results at time two, only multivariate procedures will be described here.

Either of the analytic procedures outlined above lose statistical power with small N's -- particularly the multivariate analyses. Consequently, it was deemed advantageous to pool subgroups into larger units. Four groups were formed on the basis of similar FEHR exposure: (1) no experience (O_{21}), (2) one first-semester experience (O_{22} and O_{23}), (3) an intensive second-semester experience following either no experience or a restricted experience during the first semester (O_{31a}, O_{31b}, O_{32}), and (4) an intensive experience following an extensive first-semester experience or two intensive second-semester experiences (O_{33a}, O_{33b}, O_{33c}, O_{33d}, O_{33e}).

It was recognized that in pooling the new C656 students with groups (3) and (4) we were assuming that their prior experience was at least equivalent to that of the control group from C655. However, this did not seem an untenable assumption.

The steps followed in analyzing each variable are described below:

SARES Scores. Since the MC, MI, and MR scores were commensurable, the multivariate procedure of choice was a profile analysis. As before, a profile analysis in which the hypothesis of parallel treatment profiles was not rejected was followed by a univariate analysis of the sum scores (MC + MI + MR) with planned orthogonal contrasts to test for linearity and for systematic differences between the control group and the averaged experimental groups. Significant overall tests were to be followed by three univariate analyses of covariance with orthogonal contrasts to test for linear trend and for systematic differences between control and experimental groups.

PRS Scores. The six scores which were derived from elements of final reports which had been assigned to all FEHR groups (IP, CF, M, RC, GE) were analyzed by a multivariate analysis of variance followed by orthogonal contrasts to test for linear and non-linear trends for three variable combinations: (1) all variables equally weighted, (2) equal weights on the variables common to both proposals and final reports (IP, CF, M, GE), and (3) unit weight on the method variable but zero weight on the remaining variables. These contrasts constitute tests within the multivariate model of the last three hypotheses.

Results

The results are presented in two sections, as specified above. First we will analyze time two data to determine the short term effects of FEHR, and second we will analyze data over the entire experiment. Within each section, results are presented in order of the hypotheses with which they are associated.

Results for O_2. Analysis. The time two results are organized into two parts: a univariate analysis of covariance (ANCOVA) of the FT scores, and a multivariate analysis of the SARES scores.

1. ANCOVA of FT. The summary of the analysis of covariance of the FT scores appears in table 4.1. It was observed that both the original means and the adjusted means increased monotomically with increased exposure to FEHR, although the differences among adjusted means were somewhat smaller. The test for a linear trend among the adjusted means was quite significant ($p = .0228$), but the test for a non-linear trend was not ($p = .1458$). The results were interpreted as unequivocal support for the hypothesized monotonic increase in achievement with increasing exposure to FEHR.

TABLE 4.1. ANALYSIS OF COVARIANCE OF FT SCORES USING E1 AS COVARIATE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	1672.7	836.36	6.854	.002
Zero Slope	1	2639.5	2639.5	21.631	.000
Error	49	5979.3	122.03		
Equality of Means (w/o Covariates)	2	2330.3	1165.2	6.759	.002
Error	50	8618.8	172.38		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
E1	1.103	.237	4.650	.000

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)
Mean	56.600	68.545	73.333
Adj. Mean	57.470	68.600	69.508
(Std. Error)	2.217	2.355	4.584
Intercept	16.775	27.906	28.814
Sample Size	25	22	6

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	12.039	2.350	.022
Curvelinear Trend	-10.223	-1.478	.145

2. Profile Analysis of SARES Scales. The results of the profile analysis of the MC, MI, and MR scales from the SARES instrument appear in table 4.2. It was observed that the means of the MC scores were in the predicted order of monotonic increase, but that group 1 scored higher than group 2 on both the MI and MR scales. The differences in group profiles were not found to be significant ($p > .05$), but there were significant differences among the means of variables averaged over groups ($\overline{MC} = 3.0154$, $\overline{MI} = 3.6298$, and $\overline{MR} = 3.6181$), and significant differences among groups averaged over variables ($\overline{G1} = 3.3789$, $\overline{G2} = 3.0810$, and $\overline{G3} = 3.8034$).

For reasons described previously, the effects of variables within the profile analysis were tested by separate univariate ANCOVA's of each variable. The results of these analyses appear in tables 4.3, 4.4, and 4.5. It was observed that there was a significant linear trend for and no curvelinear trend for MC scores, but that for both MI and MR scores the non-linear trend was not significant. Put another way, the group 2 mean was significantly lower than the average of the group 1 and 3 means for both these variables. Although the observed value of the group 2 mean was lower than the group 1 mean for both scales, neither of these differences was statistically significant ($p > .10$).

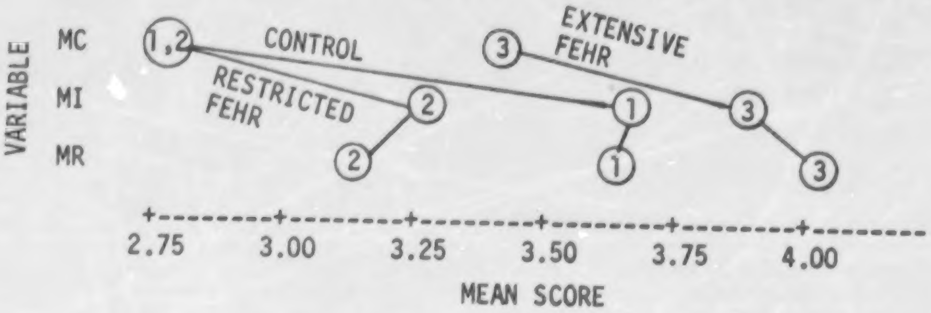
Results of Overall Analysis (Observation Times 2 and 3). The analysis of overall results is organized in two parts. First a profile analysis of the mean competency (MC), mean interest (MI), and mean relevancy (MR) scores derived from the Self Assessment of Research and Evaluation Skills (SARES) instrument was conducted. Second, a multiple analysis of variance (MANOVA) of the scales derived from part B of the product rating sheet was conducted. Since a formal review of the research, a budget, and a logistical plan were explicitly excluded from some of the FEHR assignments, the RC and LG scores were excluded from analysis. The scores analyzed were: IP (introduction and problem statement), CF (conceptual framework or theory), M (method), RC (results and conclusions), and GE (general evaluation).

TABLE 4.2. PROFILE ANALYSIS OF SARES SCORES AT TIME 2

TABLE OF MEANS

Variable	Group 1	Group 2	Group 3	Variable Means
MC	2.783	2.793	3.469	3.015
MI	3.684	3.301	3.903	3.629
MR	3.669	3.148	4.037	3.618
Group Means	3.378	3.081	3.803	3.421

GRAPHICAL DISPLAY OF PROFILES



PROFILE ANALYSIS

Tests on Groups	T-SQUARE	F-STAT.	DF	SIGNIF.
Parallelism of Profiles	Max. Root=	.096	2,-.5,23.5	NS@.05
Equality of Variable Means	61.029	29.904	2 ,49	.000
No Group Differences		3.641	2 ,50	.033

TABLE 4.3. ANALYSIS OF COVARIANCE OF MC WITH E1 AS COVARIATE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	2.976	1.488	3.973	.025
Zero Slope	1	.901	.901	2.405	.127
Error	49	18.354	.374		
Equality of Means (w/o Covariates)	2	2.476	1.236	3.210	.048
Error	50	19.255	.385		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
E1	-.020	.013	-1.551	.127

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)
Mean	2.783	2.793	3.469
Adj. Mean	2.767	2.792	3.540
(Std. Error)	.122	.130	.253
Intercept	3.519	3.544	4.292
Sample Size	25	22	6

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.772	2.723	.008
Curvelinear Trend	.723	1.887	.065

TABLE 4.4. ANALYSIS OF COVARIANCE OF MI WITH E1 AS COVARIATE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	2.636	1.318	2.864	.066
Zero Slope	1	.763	.763	1.659	.203
Error	49	22.545	.460		
Equality of Means (w/o Covariates)	2	2.852	1.426	3.059	.055
Error	50	23.308	.466		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
E1	.018	.014	1.288	.203

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)
Mean	3.684	3.301	3.968
Adj. Mean	3.699	3.302	3.903
(Std. Error)	.136	.144	.281
Intercept	3.006	2.610	3.211
Sample Size	25	22	6

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.204	.649	.518
Curvelinear Trend	.996	2.347	.023

TABLE 4.5. ANALYSIS OF COVARIANCE OF MR WITH E1 AS COVARIATE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	4.856	2.428	3.177	.050
Zero Slope	1	.892	.892	1.167	.285
Error	49	37.445	.764		
Equality of Means (w/o Covariates)	2	5.172	2.586	3.373	.042
Error	50	38.338	.766		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
E1	.020	.018	1.080	.285

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)
Mean	3.669	3.148	4.037
Adj. Mean	3.685	3.149	3.966
(Std. Error)	.175	.186	.362
Intercept	2.936	2.400	3.218
Sample Size	25	22	6

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.281	.694	.490
Curvelinear Trend	1.354	2.474	.016

The results of the two analyses are presented in the order mentioned.

1. Profile Analysis. The results of the profile analysis of the MC, MI, and MR scores appears in table 4.6. The profiles of the four groups formed by pooling students with similar laboratory experience were compared. Group 1 had no FEHR experience, group 2 had one first-term experience, group 3 had one intensive second-term experience, and group 4 had two or more intensive experiences in the two-term interval.

   It was observed that the group means for the MC scores increased in the order 1 2 3 4, and the MI scores in the order 2 1 3 4, and the MR scores in the order 2 3 1 4. Since there was a significant departure from parallel profiles, analyses of variance with planned tests for linear and curvelinear trends were conducted separately for each variable. The results of these analyses appear in tables 4.7, 4.8, and 4.9. It was observed that there was both significant linearity and curvelinearity for the MC scores (the latter being due to the 3-4 reversal). However, post-hoc comparisons failed to produce significant differences between the MC means for groups 3 and 4 (p = .5395), the MI means of groups 1 and2 (p = .1774), the MRmeans for groups 1 and 2, or 1 and 3 (p = .1413 and .8270 respectively). For the MI and MR scores, there was not significant linearity but there was significant curvelinearity.

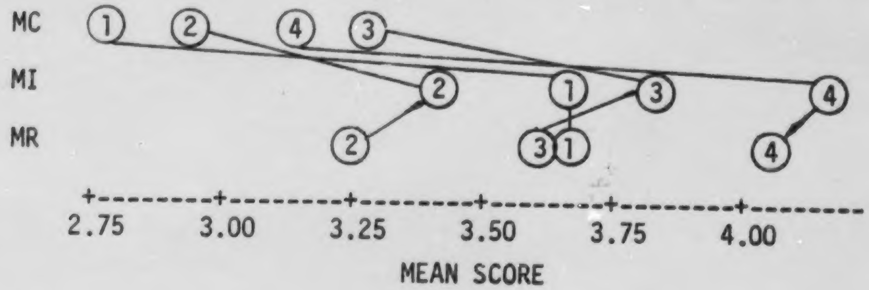
2. MANOVA of PRS Scales. The IP, CF, M, RC, and GE scales were analyzed by multivariate analysis of variance. The first step was to determine whether the new C656 students could legitimately be pooled with groups 3 and 4. To answer this question, a multivariate analysis of variance was run to compare the scores of groups 2, 3, and 4 from the original C655 students (group 1 did not appear because the control groups completed no projects) with those of new subjects in groups 5 and 6. The results of the MANOVA appear in table 4.10.

TABLE 4.6. PROFILE ANALYSIS OF SARES SCORES FOR GROUPS POOLED OVER TIMES 2 AND 3

TABLE OF MEANS

Variable	Group 1	Group 2	Group 3	Group 4	Variable Means
MC	2.783	2.962	3.277	3.157	3.045
MI	3.684	3.439	3.826	4.180	3.782
MR	3.669	3.325	3.620	4.066	3.670
Group Means	3.378	3.242	3.574	3.801	3.499

GRAPHICAL DISPLAY OF PROFILES



PROFILE ANALYSIS

Tests on Groups	T-SQUARE	F-STAT.	DF	SIGNIF.
Parallelism of Profiles	Max. Root=	.139	2,0.,46.5	.023
Equality of Variable Means	127.48	63.074	2 ,95	.000
No Group Differences		3.321	3 ,96	.023

TABLE 4.7. UNIVARIATE ANALYSIS OF VARIANCE OF MC

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	3	3.787	1.262	3.006	.034
Within	96	40.315	.419		
Total	99	44.103			

Equality of Variances: DF = 3,14778. F = .175 Signif. = .912

EXPOSURE LEVEL	N	MEAN	VARIANCE	STD. DEV.
1	25	2.783	.417	.646
2	27	2.962	.420	.648
3	31	3.277	.465	.682
4	17	3.157	.335	.579

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	1.047	2.313	.022
Curvelinear Trend	1.437	2.265	.025

TABLE 4.8. UNIVARIATE ANALYSIS OF VARIANCE OF MI

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	3	6.044	2.014	4.781	.003
Within	96	40.456	.421		
Total	99	46.501			

Equality of Variances: DF = 3,14778. F = 1.565 Signif. = .195

EXPOSURE LEVEL	N	MEAN	VARIANCE	STD. DEV.
1	25	3.684	.505	.711
2	27	3.439	.509	.713
3	31	3.826	.399	.632
4	17	4.180	.193	.439

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.393	.867	.387
Curvelinear Trend	1.876	2.951	.004

TABLE 4.9. UNIVARIATE ANALYSIS OF VARIANCE OF MR

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	3	5.773	1.924	2.766	.046
Within	96	66.795	.695		
Total	99	72.568			

Equality of Variances: DF = 3.14778. F = .994 Signif. = .394

EXPOSURE LEVEL	N	MEAN	VARIANCE	STD. DEV.
1	25	3.669	.876	.936
2	27	3.325	.803	.896
3	31	3.620	.606	.778
4	17	4.066	.416	.645

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.004	.007	.994
Curvelinear Trend	1.485	1.818	.072

TABLE 4.10. RESULTS OF THE MULTIVARIATE COMPARISONS OF PRODUCT RATINGS FOR ORIGINAL C655-C656 AND NEW C656 SUBJECT GROUPINGS

MULTIVARIATE ONE-WAY ANALYSIS OF VARIANCE: OVERALL TEST

Equality of Group Means: DF = 20,269.60 F = 3.599 SIGNIF. = .000

Alt. Test of Equality of Group Means: Max. Root = .586 SIGNIF. = .000

TABLE OF MEANS

	Original C655-C656 Subjects			New C656 Subjects	
Group FEHR Exp.	(2) One: C655	(3) One: C656	(4) Two: C656	(5) One: C656	(6) Two: C656
IP	9.909	11.032	13.118	10.364	13.444
CF	8.590	9.354	10.176	8.181	12.778
M	24.364	28.323	28.529	26.818	27.889
RC	23.636	26.258	30.294	28.909	27.556
GE	8.454	12.387	14.529	10.364	13.556
Sample Size	22	31	17	11	9

PAIRWISE ANALYSIS WITH UNIT WEIGHT ON ALL VARIABLES

COMPARISONS	OBSERVED VALUE	CRITICAL VALUES SIG. = .05	CRITICAL VALUES SIG. = .01
Group 2 vs. Group 3	-12.400	20.098	22.634
Group 2 vs. Group 4	-21.693	23.281	26.219
Group 2 vs. Group 5	- 9.681	26.623	29.983
Group 2 vs. Group 6	-20.268	28.527	32.127
Group 3 vs. Group 4	- 9.292	21.758	24.504
Group 3 vs. Group 5	2.718	25.302	28.495
Group 3 vs. Group 6	- 7.867	27.299	30.743
Group 4 vs. Group 5	12.011	27.898	31.418
Group 4 vs. Group 6	1.424	29.720	33.470
Group 5 vs. Group 6	10.586	32.405	36.494

TABLE 4.11. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF PRODUCT RATINGS FOR POOLED GROUPS

Equality of Group Means: DF = 10,166.00 F = 5.725 SIGNIF. = .000

Alt. Test of Equality of Group Means: Max. Root = .528 SIGNIF. = .000

TABLE OF MEANS

Exposure Level	(2)	(3)	(4)
IP	9.909	10.857	13.231
CF	8.590	9.047	11.077
M	24.364	27.929	28.308
RC	23.636	26.952	29.346
GE	8.454	11.857	14.192
Sample Size	22	42	26

Analysis Of Trends For The Final Report Combination: Unit Weight On Each Variable

COMPARISON TESTED	OBSERVED VALUE	CRITICAL VALUES	
		SIG. = .05	SIG. = .01
Linear Trend	21.199	17.589	20.275
Curvelinear Trend	- 2.177	25.699	29.624

Analysis Of Trends For The Proposal Combination: Unit Weight On All Variables But RC, Which Was Given Zero Weight

COMPARISON TESTED	OBSERVED VALUE	CRITICAL VALUES	
		SIG. = .05	SIG. = .01
Linear Trend	15.490	12.107	13.956
Curvelinear Trend	- 1.255	17.691	20.392
Linear Trend Comparison By Variable			
IP	3.321	2.559	2.950
CF	2.486	5.101	5.880
M	3.944	5.018	5.785
RC	5.709	7.948	9.762
GE	5.737	3.570	4.115
Curvelinear Trend Comparison By Variable			
IP	1.425	3.739	4.310
CF	1.572	7.454	8.592
M	- 3.185	7.333	8.453
RC	- .922	11.614	13.387
GE	- 1.067	5.217	6.013

It was observed that the prospective pooling pairs (groups 3 and 5 and groups 4 and 6) were quite similar. With unit weights on each of the five variables the observed value of T^2 for each of these comparisons was less than one-tenth the critical value for significance at the .05 level. It was concluded that the new students in groups 5 and 6 could legitimately be pooled with groups 3 and 4 respectively.

The multivariate analysis of variance of the product ratings for the pooled subject groupings appears in table 4.11. It was observed that for all variables there was a uniform increase in the mean with an increase in FEHR experience. Subsequent tests for linear and curvelinear trends were performed for a final report combination using unit weights on all variables and a proposal combination which used unit weights for all proposal variables (i.e., IP, CF, M, and GE) and zero weights for all other variables. In addition, tests for trend were run on each marginal; that is, for each variable separately. It was observed that the linear tests reached significance for the IP and GE variables, and yielded substantial positive values for all other variables. The T^2 values for curvelinear tests were generally smaller, and all were non-significant. It was concluded that there was a uniform linear increase in all five project rating scores with increased exposure to FEHR-PRACTICUM.

Summary of Experimental Evaluation Results

The data from the experimental evaluation uniformly supported the hypothesized monotonic increase with increased FEHR experience for achievement (FE scores), and project ratings (IP, CF, M,RC, and GE scores. The results for perceived achievement agreed with this pattern except for a small but insignificant decrease from exposure level 3 to level 4.

A curvelinear relationship was found for both interest in research (MI scores), and the perceived relevance or importance of research (MR scores). For these scores (MI and MR), a short or restricted exposure tended to produce somewhat less interest than the control condition. But with increased exposure the scores equalled, then exceeded those of the controls. Although the patterns were similar, this tendency was considerably stronger for interest (MI) than for relevancy (MR) scores.

STUDY TWO: FIELD TRIALS

The data from the experimental evaluation (study one) presented previously was merged with data from the field trials at the University of Michigan-Flint College, Indiana University, Michigan State University, Ohio State University, and Western Michigan University to form the data base for study two.

Purpose. The specific purpose of study two was to determine whether the monotonic increases in achievement and interest which were found in the relatively controlled conditions of the experimental evaluation could be generalized to sites other than the University of Michigan and for purposes other than the "stat-lab" role investigated in study one. In particular, we wished to answer these critical questions:

(1) Is there a systematic relationship between the degree of exposure to FEHR and the following: (a) project quality, (b) perceived achievement, and (c) attitude towards research?

(2) Are all problems equally effective?

(3) How do the three class types (general research methods, research methods for specialized content, and in-service workshops) compare in terms of project quality, perceived achievement, and research attitude?

(4) Is FEHR more effective for teaching general research methods when it is integrated into course content than when it is used as an independent laboratory experience.

Method

Subjects. The subjects for the field evaluation consisted of students from all twenty classes described in the general description at the beginning of chapter IV. For purpose of this overall evaluation there was no random sampling: all classes were sampled as intact groups. It was assumed that these classes were reasonably representative of both full-time and part-time (in service) students in graduate education.

Treatments. The treatments to be compared consisted of the four classroom classification dimensions listed under treatments in figure 4.9. The full title for each dimension is listed below in order of tabular appearance, with the keyword from the table heading underlined: (1) degree of exposure to FEHR, (2) problem content, (3) type of class, and (4) degree of integration between class content and the FEHR project assigned. Detailed descriptions of each class and the rationale behind each classification system were given in the subjects section at the beginning of chapter IV. A brief summary of each treatment classification is provided in context with the design.

Design. The four treatment dimensions were the only factors to be studied in the field evaluation of FEHR. The factors and the number of levels in each were: exposure, with five levels; problem, with eight levels; type, with three levels; and integration, with two levels. Obviously, the 240 cells needed for a complete factorial design were beyond the scope of the study. Because of the breadth of choice required by off-campus users as a condition of participation, even a balanced incomplete-blocks design became impossible. The alternative strategy was to treat each dimension as a separate intact-groups experiment, and attempt to control statistically for variations attributable to the other factors was also rejected. Because of many empty datum cells, it was impossible to use analysis of covariance for this purpose.

As a result of these restrictions, the following compromise strategy was developed. Wherever there was sufficient redundancy to permit it, the evaluation of a factor would be conducted at a single level of the other factors. The means from these single level evaluations could then be used as covariates in the analysis of factors for which single level evaluation was not possible. Since this strategy, in effect, creates four different designs each necessitating a different analytic plan, the remainder of this section is

		TREATMENT VARIABLES				DEPENDENT VARIABLES			
						PROJECT RATINGS		GOAL RATINGS	ORS QUEST.
CLASS	N	EXPOSURE LEVEL	PROBLEM AREA	TYPE OF CLASS	INTEGRATION LEVEL	PROP.	F.R.	(GAQ)	(ORSQ)
Experimental Evaluation									
1. C655 (UM) Sec. 1, F 72	25	1	None	Gen. Res.	Incomplete	X	X		
2. C655 (UM) Sec. 2, F 72	22	3	Choice 1-8	Gen. Res.	Incomplete	X	X		
3. C655, Sec. 2 & K680, F 72	6	4	Choice 1-8	Gen. Res.	Incomplete	X	X		
4. C656, W 73: (a) S's from (1)	15	3	Choice 1-8	Gen. Res.	Incomplete	X	X	X	X
(b) New S's	18	3	Choice 1-8	Gen. Res.	Incomplete	X	X	X	X
5. C656, S's from (2), W 73	21	4	Choice 1-8	Gen. Res.	Incomplete	X	X	X	X
6. C656, S's from (3), W 73	6	5	Choice 1-8	Gen. Res.	Incomplete	X	X	X	X
Field Trials									
7. Mich. State U., W 73	19	3	REMAR	Gen. Res.	Incomplete	X	X		X
8. Ohio State U., W 73	13	4	REMAR	Gen. Res.	Incomplete	X	X		
9. West. Mich. U., W 73	5	3	REMAR	Gen. Res.	Incomplete	X	X		
10. Indiana U., W 73	8	4	PEP	Spec. Ed. Res.	Incomplete	X	X	X	X
11. C699 (UM), F 72	6	4	Choice 1-8	Prog.Eval. Wkshp.	Incomplete	X	X	X	X
12. C699 Flint, W 73	7	3	REMAR	Prog.Eval. Wkshp.	Incomplete	X	X	X	X
13. C699 (UM), W 73	6	4	Choice 1-8	Prog.Eval. Wkshp.	Complete	X	X	X	X
14. C655 (um) Sec. 1, F 73	46	4	REMAR	Gen. Res.	Complete	X	X	X	X
15. C655 (UM) Sec. 2, F 73	30	4	REMAR	Gen. Res.	Complete	X	X	X	X
16. C656 (UM) Sec. 1, W 74	32	5	TQUEST	Gen. Res.	Complete	X		X	X
17. C656 (UM) Sec. 2, (a) W 74	19	5	TQUEST	Gen. Res.	Complete			X	X
(b) New S's	4	4	TQUEST	Gen. Res.	Complete			X	X
FIELD TEST TOTAL N	308								
Control Groups: Field Trials									
18. A705 (UM), W 73	30	1	---	Phil. of Educ.	---				X
19. C607/C654, W 73	11	2	---	Theory Msrmnt.	---				X
20. Ph.D. Proposals (sampled randomly from Ed. Psych. files).	9	2	---	Ind. Res.	---	X			
GRAND TOTAL N	358								

Figure 4.9. Data Classification for the Study Two Field Evaluation.

organized by factors. For each factor we will present: (1) a brief description of the various levels of the factor, (2) a summary of the data to be analyzed, (3) specific hypotheses, (4) an analytic plan, (5) the results of the analysis, and (6) a brief summary statement.

Factor One: Exposure Level

Description. The degree of exposure in figure 4.9 combines the complexity/difficulty (pattern of FEHR tasks assigned) with the number of exposures, and the total duration of exposures to form pooled groups similar to those used in the experimental evaluation. However, in this study there were two control groups: subjects who had no training in research design and statistics, and subjects who had taken a statistics course other than those involved in FEHR. The five exposure levels, in ascending order of experience with FEHR were:

(1) Subjects in this level had no experience with FEHR or a research design/statistics course.

(2) Subjects in this level had no experience with FEHR, but at least one experience in a research design/statistics course which was not associated with FEHR or members of the FEHR-PRACTICUM project.

(3) Subjects in this level had experienced a one-semester FEHR project -- usually somewhat restricted. It frequently required a proposal, but usually without a budget or funds negotiations. A final report was required. It usually consisted of a problem statement, a simplistic conceptual framework, a fairly complete method section, a summary of results and conclusions, and a recommended decision with supporting rationale. Subjects from the restricted level of the experimental evaluation study were classified in this level for purposes of the field study. Most subjects in this level had no prior statistical training.

(4) Subjects in this level had experienced a one-semester intensive FEHR project requiring both a formal proposal and a final report. Both documents were comprehensive in

coverage, but there was a limited review of the literature section -- usually only the studies provided in the information bank were assigned. It is noteworthy, however, that some students in these classes went well beyond the requirements, and one or two completed very comprehensive reviews. Both the extensive and intensive treatment levels from the experimental evaluation study were included in this level of the field evaluation. About half of the subjects in this level had one semester of prior statistical training.

(5) Subjects in this level had experienced either two or more FEHR projects, usually (but not always) over a two-semester period, or else one fully integrated FEHR project at the first-semester level (i.e., to statistical novices). The only fully integrated projects in the study occurred in the C655 class conducted during the fall semester 1973 at the University of Michigan. For subjects completing two projects, at least one must have been conducted at the intensive level. Subjects from the extensive/intensive level of the experimental evaluation were included in this level of the field study. About half the subjects at this level had one semester of prior statistical training, and half had no prior training.

Data Matrix. The data sets to be analyzed across exposure groups are summarized in figure 4.10. A key to the meaning of each symbol appears in the bottom cell of the figure.

Missing data sets occurred wherever the measures were undefined for the group concerned. Obviously, proposals were not available for control subjects with no research experience (group 1). However, proposals were available for the small subset of the experienced controls (group 2) who had written a dissertation proposal. On the other hand, final reports were not available in either control group. Similarly, the GA ratings for groups 1 and 2 are missing because the goals being rated were irrelevant for the control classes.

SAMPLING PROCEDURE	EXPOSURE LEVEL	CONTROL VARIABLES	PROPOSAL SCORES	FINAL REPORT SUM	GOAL ATTAINMENT	INTEREST RATINGS	DIFFICULTY RATINGS
IG	1	missing	missing	missing	missing	I_1	D_1
IG	2	missing	P_2	missing	missing	I_2	D_2
IG	3	C_3	P_3	F_3	GA_3	I_3	D_3
IG	4	C_4	P_4	F_4	GA_4	I_4	D_4
IG	5	C_5	P_5	F_5	GA_5	I_5	D_5

Key to Symbols Used Above

Symbol	Description
IG	Indicates that sampling was by intact groups.
C	The set of control variable scores for a group. There are two scores for each individual: the mean of his class type and the mean of his integration classification.
P	The set of ratings on commonly assigned proposal tasks (IP, CF, M, GE) for the group.
F	The sum of ratings on commonly assigned final report tasks (IP + CF + M + GE) for the group.
GA	The set of goal attainment ratings (goal assessment questionnaire) for each group.
I	The set of interest ratings from the ORS questionnaire (ICR, IEV, IPW, IRP) for the group.
D	The set of difficulty ratings from the ORS questionnaire (DCR, DEV, DPW, DRP) for the group.

Figure 4.10. Diagramatic Summary of the Data Matrix Stratified by Exposure Level.

Hypotheses. The hypotheses to be tested are listed in order of the variable listing from left to right in figure 4.10.

(1) $P_2 < P_3 < P_4 < P_5$

(2) $F_3 < F_4 < F_5$

(3) $GA_3 < GA_4 < GA_5$

(4) $I_1 < I_2 < I_3 < I_4 < I_5$

(5) $D_1 < D_2 < D_3 < D_4 < D_5$

A brief summary statement which attempts to integrate and/or reconcile the multivariate and univariate results is provided for each variable set.

Results for Project Report Variables. The results of the multivariate analysis of variance of the proposal variables appear in table 4.12. It was observed that the linear trend was significant for the entire variable set and for each of the individual variables except IP, and even there the observed contrast value was very near to the critical value at the .05 level of significance.

The results of the univariate analyses of covariance for each variable individually appear in tables 4.13 to 4.16. It was observed that a linear trend was maintained for each variable except the M ratings, in which the adjusted mean level 4 mean was slightly higher than that of level 5.

The analysis of covariance for the PC ratings appears in table 4.17. A significant linear trend was observed, and a non-significant curvelinear trend. The analysis of covariance for the FC scores appears in table 4.18. Again. a highly significant linear trend (p = .004) and a non-significant curvelinear trend were observed. The pattern of mean scores for the composite final report ratings was entirely consistent with the pattern for proposal ratings.

It was concluded that the overall pattern of results supported the hypothesis there was a monotonic increase in proposal ratings with increased exposure to FEHR. The results also indicate that subjects given two or more exposures to FEHR obtained proposal ratings which were at least as high as the ratings of the dissertation proposals in control group 2.

TABLE 4.12. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF PROPOSAL RATINGS STRATIFIED BY EXPOSURE LEVEL

Equality of Group Means: DF = 12,547.96 F = 5.510 SIGNIF. = .000

Alt. Test of Equality of Group Means: Max. Root = .251 SIGNIF. = .000

TABLE OF MEANS

Exposure Level	(2)	(3)	(4)	(5)
IP	14.667	11.600	11.119	13.226
CF	11.778	8.745	8.976	14.962
M	14.778	24.191	26.000	30.755
GE	10.667	10.618	10.810	13.962
Sample Size	9	110	42	53

	OBSERVED VALUE	CRITICAL VALUES	
		SIG. = .05	SIG. = .01
Comparisons For The Entire Set With Unit Weights On Each Variable			
Control vs. Experimental (Level 2 vs. 3, 4, & 5)	29.298	60.906	69.265
Linear Trend (Levels 3, 4, & 5)	17.751	9.930	11.293
Curvelinear Trend (Levels 3, 4, & 5)	14.251	20.845	23.706
Control vs. Experimental Comparison By Variable			
IP	- 8.054	12.638	14.373
CF	- 2.649	22.608	25.711
M	36.612	25.568	29.077
GE	3.390	16.779	19.082
Linear Trend Comparison By Variable (Group 2 Omitted)			
IP	1.626	2.060	2.343
CF	6.216	3.686	4.192
M	6.563	4.168	4.740
GE	3.344	2.735	3.111
Curvelinear Trend By Variable (Group 2 Omitted)			
IP	2.588	4.325	4.919
CF	5.755	7.737	8.799
M	2.945	8.750	9.951
GE	2.961	5.742	6.530

TABLE 4.13. ANALYSIS OF COVARIANCE OF IP RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	105.96	52.982	5.469	.004
Zero Slope	2	44.241	22.120	2.283	.104
Error	200	1937.3	9.686		
Equality of Means (w/o Covariates)	2	174.59	87.294	8.898	.000
Error	200	1981.6	9.809		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.089	.044	1.991	.047
Integration Level	.344	.506	.679	.497

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	11.614	11.119	13.435
Adj. Mean	11.680	11.196	13.276
(Std. Error)	.311	.514	.429
Intercept	5.923	5.438	7.518
Sample Size	101	42	62

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	1.595	2.958	.003
Curvelinear Trend	2.564	2.143	.033

TABLE 4.14. ANALYSIS OF COVARIANCE OF CF RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	320.57	160.29	7.162	.001
Zero Slope	2	1765.1	882.54	39.435	.000
Error	200	4475.9	22.380		
Equality of Means (w/o Covariates)	2	1335.1	667.54	21.606	.000
Error	202	6241.0	30.896		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.205	.068	3.010	.002
Integration Level	6.313	.769	8.202	.000

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	8.930	8.976	14.500
Adj. Mean	9.375	11.128	12.317
(Std. Error)	.473	.781	.652
Intercept	-11.744	-9.991	-8.802
Sample Size	101	42	62

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	2.941	3.588	.000
Curvelinear Trend	- .563	- .310	.756

TABLE 4.15. ANALYSIS OF COVARIANCE OF M RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	478.57	239.29	6.327	.002
Zero Slope	2	2455.0	1227.5	32.459	.000
Error	200	7563.6	37.818		
Equality of Means (w/o Covariates)	2	787.41	393.71	7.938	.000
Error	202	10019.	49.597		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.230	.088	2.601	.010
Integration Level	7.497	1.000	7.492	.000

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	23.921	26.000	28.435
Adj. Mean	24.442	28.563	25.850
(Std. Error)	.615	1.016	.847
Intercept	.136	4.256	1.544
Sample Size	101	42	62

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	1.408	1.321	.187
Curvelinear Trend	-6.832	-2.890	.004

TABLE 4.16. ANALYSIS OF COVARIANCE OF GE RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	102.43	51.216	3.910	.021
Zero Slope	2	846.21	423.11	32.302	.000
Error	200	2619.7	13.098		
Equality of Means (w/o Covariates)	2	387.68	193.84	11.297	.000
Error	202	3465.9	17.158		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.325	.052	6.227	.000
Integration Level	2.813	.588	4.778	.000

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	10.386	10.810	13.484
Adj. Mean	10.710	11.647	12.389
(Std. Error)	.362	.597	.498
Intercept	-12.459	-11.521	-10.779
Sample Size	101	42	62

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	1.679	2.678	.008
Curvelinear Trend	- .195	- .140	.888

TABLE 4.17. ANALYSIS OF COVARIANCE OF THE PC RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	2527.5	1263.7	7.440	.000
Zero Slope	2	14975.	7487.5	44.082	.000
Error	200	33971.	169.85		
Equality of Means (w/o Covariates)	2	9093.5	4546.7	18.764	.000
Error	202	48946.	242.31		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.851	.188	4.525	.000
Integration Level	16.968	2.120	8.001	.000

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	54.851	56.905	69.855
Adj. Mean	56.207	62.534	63.833
(Std. Error)	1.305	2.153	1.796
Intercept	-18.144	-11.818	-10.519
Sample Size	101	42	62

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	7.625	3.376	.000
Curvelinear Trend	-5.027	-1.003	.316

TABLE 4.18. ANALYSIS OF COVARIANCE OF THE FC RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	7914.7	3957.4	11.800	.000
Zero Slope	2	9312.5	4656.3	13.884	.000
Error	126	42256.	335.37		
Equality of Means (w/o Covariates)	2	7497.8	3748.9	9.305	.000
Error	128	51569.	402.88		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.682	.326	2.088	.038
Integration Level	15.730	4.907	3.205	.001

Exposure Level			
Mean	73.747	84.220	98.545
Adj. Mean	72.933	86.725	95.052
(Std. Error)	2.071	2.900	5.734
Intercept	.867	14.660	22.987
Sample Size	79	41	11

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	22.119	3.617	.000
Curvelinear Trend	- 5.465	-.641	.522

Results for Goal Attainment. The results of the multivariate analysis of variance of the goal attainment ratings are summarized in table 4.19. It was observed that there was no consistent linear or curvelinear trends across goals. In the overall tests for the entire variable set (with unit weights on each rating) both the linear and the curvelinear contrasts yielded observed values well below the critical value for significance at the .05 level. In addition, neither of the marginal contrasts reached significance for any of the goal attainment ratings.

Because there was neither a persistent pattern nor a significant trend for any single variable, analyses of covariance of the attainment ratings for the individual goals were not conducted; instead, we went directly to an analysis of the mean attainment rating averaged over the nine goals.

The analysis of covariance of the mean ratings appears in table 4.20. It was observed that the mean ratings for level four subjects was lower than that for either level three or five. Nevertheless, the observed contrast value for linear trend was considerable ($p = .0555$). The curvelinear trend was, of course, highly significant. Post hoc comparisons of level three with the other two levels yielded insignificant t ratios ($p > .20$). It was concluded that the results of the GA analyses were inconclusive with respect to the hypothesized monotonic increase in goal attainment with increased exposure to FEHR: they neither support nor deny the hypothesis.

Results for Interest Variables. The results of the multivariate analysis of variance of the interest ratings are summarized in table 4.21. It was observed that there was a uniform increase in the observed means of exposure levels two through five for interest ratings on classical research methods (ICR), program evaluation (IPE), and proposal writing (IPW). For the last variable -- the research practicum (IRP) -- there was an increase for levels three through five, but level two was somewhat higher than level three. For classical research and program evaluation, the mean for the control group with no research experience (level 1) was higher than the mean for the control group with some statistical training (level 2) and about the

TABLE 4.19. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF THE GOAL ATTAINMENT RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS EXCLUDED)

Equality of Group Means: DF = 18,106.00 F = 2.589 Sig. = .0013

Alt. Test of Equality of Group Means: Max. Root = .6508 Sig. = .0041

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Goal 1	4.075	3.513	4.100
Goal 2	3.530	3.222	3.944
Goal 3	3.651	3.444	4.177
Goal 4	3.560	3.305	4.177
Goal 5	3.818	3.083	4.100
Goal 6	3.530	3.805	4.033
Goal 7	3.424	3.416	3.955
Goal 8	3.765	3.277	4.077
Goal 9	3.878	3.666	4.216
Sample Size	22	12	30

	OBSERVED VALUE	CRITICAL VALUES	
		SIG. = .05	SIG. = .01
Comparisons For The Entire Set With Unit Weights On Each Variable			
Linear Trend	3.548	5.757	6.541
Curvelinear Trend	8.546	13.167	14.960
Linear Trend Comparison By Variable			
Goal 1	.024	1.110	1.261
Goal 2	.414	1.081	1.229
Goal 3	.526	1.016	1.154
Goal 4	.617	1.122	1.275
Goal 5	.281	.966	1.098
Goal 6	.503	.972	1.105
Goal 7	.531	.970	1.103
Goal 8	.312	.895	1.017
Goal 9	.337	.915	1.040
Curvelinear Trend Comparison By Variable			
Goal 1	1.148	2.538	2.884
Goal 2	1.030	2.474	2.811
Goal 3	.940	2.324	2.640
Goal 4	1.127	2.568	2.917
Goal 5	1.751	2.210	2.511
Goal 6	.047	2.224	2.527
Goal 7	.546	2.220	2.522
Goal 8	1.287	2.048	2.327
Goal 9	.762	2.094	2.379

TABLE 4.20. ANALYSIS OF COVARIANCE OF THE ATTAINMENT RATING AVERAGED OVER GOALS AND STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	.778	.389	5.025	.008
Zero Slope	2	.720	.360	4.651	.011
Error	113	8.755	.077		
Equality of Means (w/o Covariates)	2	2.082	1.041	12.635	.000
Error	115	9.476	.082		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.847	.426	1.985	.049
Integration Level	- .010	.122	- .082	.934

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.537	3.389	3.793
Adj. Mean	3.547	3.368	3.779
(Std. Error)	.045	.083	.087
Intercept	.513	.334	.745
Sample Size	72	12	34

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.231	1.935	.055
Curvelinear Trend	.590	3.100	.002

TABLE 4.21. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF THE INTEREST RATINGS STRATIFIED BY EXPOSURE LEVEL

Equality of Group Means: DF = 16,645.25 F = 5.229 Sig. = .000

Alt. Test of Equality of Group Means: Max. Root = .189 Sig. = .000

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)	(4)	(5)
ICR	4.193	3.978	4.193	4.541	4.687
IPE	4.139	4.155	4.279	4.623	4.917
IPW	4.634	3.855	4.184	4.299	4.526
IRP	4.839	4.408	3.987	4.628	5.011
Sample Size	22	26	94	33	44

	OBSERVED VALUE	CRITICAL VALUES SIG. = .05	CRITICAL VALUES SIG. = .01
Comparisons For The Entire Set With Unit Weights On Each Variable			
Control vs. Experimental	5.142	14.034	15.745
Linear Trend	9.678	10.741	12.051
Curvelinear Trend	.804	4.462	5.007
Control vs. Experimental Comparison By Variable			
ICR	2.329	3.903	4.378
IPE	2.753	3.765	4.224
IPW	.549	3.687	4.137
IRP	- .490	4.767	5.349
Linear Trend Comparison By Variable			
ICR	2.474	2.987	3.351
IPE	2.629	2.881	3.233
IPW	2.126	2.822	3.166
IRP	2.448	3.649	4.094
Curvelinear Trend Comparison By Variable			
ICR	- .068	1.241	1.392
IPE	.171	1.197	1.343
IPW	- .101	1.172	1.315
IRP	.803	1.516	1.701

same as the minimum FEHR group (level 3). On the other hand, the mean ratings for proposal writing and research practicum for the no experience group were higher than the mean ratings for the group with maximum FEHR experience (level 5). Despite the relatively uniform pattern, none of the three planned contrasts (control vs. experimental, linearity within the experienced groups 2 to 5, or curvelinearity within the experienced groups) was significant for the entire set of variables with unit weight on each rating, or for the marginal (variable-by-variable) comparisons. Nevertheless, consistently high positive values were observed for the underlying monotonic trend. Consequently, univariate analyses were computed for each interest rating.

An analysis of covariance was computed for each interest rating stratified by exposure (levels 3 to 5 only) and covaried on class type and level of integration with course content. The results of these analyses appear in tables 4.22 through 4.25. It was observed that the linear comparison among adjusted means was highly significant for all three variables, and the curvelinear comparison was not significant. In addition, the analysis of covariance of the means of the four interest ratings (table 4.26) produced a highly significant linear trend ($p = .004$) but a nonsignificant curvelinear trend.

It was concluded that within the experimental groups (levels 3 to 5) the overall trend of the results provided clear support for the hypothesized monotonic increase in interest with increased exposure to FEHR. The meaning of the relationship between the control and experimental groupings is unclear on the basis of the interest ratings by themselves.

Results for Difficulty Variables. The results of the multivariate analysis of variance of the difficulty ratings appear in table 4.27. It was observed that the pattern of difficulty means was similar to the pattern for interest, but the linearity within the experimental FEHR groups (levels 3 to 5) was not quite so consistent. As compared to interest ratings, the observed contrast values for the control/experimental comparison were slightly closer to significance, and the linear comparisons slightly further from

TABLE 4.22. ANALYSIS OF COVARIANCE OF THE ICR RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	6.979	3.489	4.691	.010
Zero Slope	2	18.296	9.147	12.298	.000
Error	166	123.47	.743		
Equality of Means (w/o Covariates)	2	8.276	4.138	4.904	.008
Error	168	141.77	.843		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.150	.458	2.508	.013
Integration Level	.258	.317	.813	.417

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	4.193	4.541	4.687
Adj. Mean	4.172	4.693	4.618
(Std. Error)	.114	.166	.241
Intercept	-1.213	- .692	- .767
Sample Size	94	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.445	1.403	.162
Curvelinear Trend	- .595	-1.368	.173

TABLE 4.23. ANALYSIS OF COVARIANCE OF THE IPE RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	10.855	5.427	7.882	.000
Zero Slope	2	13.258	6.628	9.627	.000
Error	166	114.30	.688		
Equality of Means (w/o Covariates)	2	12.742	6.370	8.390	.000
Error	168	127.55	.759		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.464	.441	3.317	.001
Integration Level	- .216	.305	- .708	.479

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	4.279	4.623	4.917
Adj. Mean	4.164	4.657	5.137
(Std. Error)	.110	.160	.232
Intercept	-1.969	-1.475	- .996
Sample Size	94	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.972	3.185	.001
Curvelinear Trend	- .136	- .032	.974

TABLE 4.24. ANALYSIS OF COVARIANCE OF THE IPW RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	4.108	2.054	3.669	.027
Zero Slope	2	10.305	5.152	9.205	.000
Error	166	92.922	.559		
Equality of Means (w/o Covariates)	2	3.499	1.749	2.847	.060
Error	168	103.23	.614		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.328	.398	3.337	.001
Integration Level	- .230	.275	- .836	.404

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	4.184	4.299	4.526
Adj. Mean	4.075	4.320	4.744
(Std. Error)	.099	.144	.209
Intercept	-1.443	-1.198	- .773
Sample Size	94	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.669	2.432	.016
Curvelinear Trend	.178	.473	.636

TABLE 4.25. ANALYSIS OF COVARIANCE OF THE IRP RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	23.260	11.630	9.991	.000
Zero Slope	2	25.843	12.922	11.101	.000
Error	166	193.22	1.164		
Equality of Means (w/o Covariates)	2	34.068	17.034	13.064	.000
Error	168	219.06	1.303		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.645	.573	2.866	.004
Integration Level	.076	.397	.192	.847

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.987	4.628	5.011
Adj. Mean	3.909	4.760	5.077
(Std. Error)	.143	.208	.302
Intercept	-3.404	-2.553	-2.237
Sample Size	94	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	1.167	2.939	.003
Curvelinear Trend	- .534	- .981	.327

TABLE 4.26. ANALYSIS OF COVARIANCE OF THE MEAN OF ALL INTEREST RATINGS

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	10.004	5.002	8.242	.000
Zero Slope	2	16.289	8.144	13.422	.000
Error	169	102.56	.606		
Equality of Means (w/o Covariates)	2	12.953	6.476	9.318	.000
Error	171	118.85	.695		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.401	.414	3.382	.000
Integration Level	- .026	.286	- .092	.926

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	4.152	4.523	4.785
Adj. Mean	4.075	4.607	4.892
(Std. Error)	.101	.149	.219
Intercept	-2.029	-1.497	-1.212
Sample Size	97	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.817	2.853	.004
Curvelinear Trend	- .246	- .628	.530

TABLE 4.27. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF THE DIFFICULTY RATINGS STRATIFIED BY EXPOSURE LEVEL

Equality of Group Means: DF = 16,639.14 F = 2.125 Sig. = .006

Alt. Test of Equality of Group Means: Max. Root = .091 Sig. = .036

TABLE OF MEANS

Exposure Level	(1)	(2)	(3)	(4)	(5)
DCR	3.176	3.346	3.760	3.640	3.644
DPE	3.375	3.487	3.840	3.661	3.996
DPW	3.943	3.896	3.946	3.979	4.078
DRP	4.204	3.711	4.184	3.750	4.011
Sample Size	22	26	92	33	44

	OBSERVED VALUE	CRITICAL VALUES SIG. = .05	SIG. = .01
Comparisons For The Entire Set With Unit Weights On Each Variable			
Control vs. Experimental	5.564	12.056	13.593
Linear Trend	3.168	9.224	10.401
Curvelinear Trend	- .591	3.836	4.325
Control vs. Experimental Comparison By Variable			
DCR	2.520	3.400	3.934
DPE	2.406	3.994	4.503
DPW	.492	3.559	4.012
DRP	.144	4.332	4.884
Linear Trend Comparison By Variable			
DCR	.775	2.602	2.934
DPE	1.347	3.056	3.446
DPW	.581	2.723	3.070
DRP	.464	3.315	3.737
Curvelinear Trend Comparison By Variable			
DCR	- .409	1.082	1.220
DPE	- .018	1.271	1.433
DPW	.048	1.132	1.276
DRP	- .211	1.378	1.554

significance. But again none of the contrasts within the multivariate model reached significance.

The results of the analyses of covariance of the difficulty scores stratified by exposure level within the FEHR experimental groups are summarized in tables 4.28 through 4.31. For both difficulty ratings for classical research methods (DCR) and program evaluation (DPE), the covariates have considerably altered the picture. The adjusted means now decrease as experience increases, but the change is not significant in either case. The adjusted difficulty ratings means for proposal writing (DPW) showed a slight but insignificant increase. The adjusted difficulty ratings for the research practicum (DRP) on the other hand yielded a significant curvelinear trend. However the analysis of covariance did not yield significant linear or curvelinear contrasts.

It was concluded that the overall pattern of evidence neither supported nor denied the hypothesized monotonic relationships between perceived difficulty and increased exposure to FEHR.

Factor Two: Problem Content Areas

Description. Eight different problems were tested in the field study. Since these were described fully in earlier chapters of this report, only the titles are reproduced here. The problems field tested were:

(1) The Perceptual Education Problem (PEP)
(2) The Remedial Arithmetic Problem (REMAR)
(3) The Extended School Year Problem (EXTSY)
(4) The Early Childhood Education Problem (HEADSTART)
(5) The Reading Assessment Problem (READ)
(6) The Validation of a Teacher Rating Questionnaire Problem (TQUEST)
(7) The Remedial Mathematics for Adults Problem (RMA)
(8) The Busing to Achieve Integration Problem (BUS)

Data Matrix. The data to be analyzed across problems is similar to that used for the exposure contrasts. However, when the data is stratified by problem, a number of empty cells occur -- largely because we were unable to administer the ORS and Goal Assessment

TABLE 4.28. ANALYSIS OF COVARIANCE OF THE DCR RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	.716	.358	.730	.483
Zero Slope	2	7.657	3.828	7.799	.000
Error	164	80.511	.490		
Equality of Means (w/o Covariates)	2	.577	.288	.543	.581
Error	166	88.169	.531		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.692	.532	1.299	.195
Integration Level	.360	.266	1.350	.178

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.760	3.640	3.644
Adj. Mean	3.793	3.776	3.472
(Std. Error)	.095	.135	.201
Intercept	.625	.608	.304
Sample Size	92	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	- .320	-1.204	.230
Curvelinear Trend	- .287	- .804	.422

TABLE 4. 29. ANALYSIS OF COVARIANCE OF THE DPE RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	.517	.258	.354	.701
Zero Slope	2	5.295	2.647	3.631	.028
Error	164	119.57	.729		
Equality of Means (w/o Covariates)	2	2.124	1.062	1.412	.246
Error	166	124.86	.752		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.084	.648	.130	.896
Integration Level	.520	.325	1.600	.111

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.840	3.661	3.996
Adj. Mean	3.924	3.815	3.705
(Std. Error)	.116	.165	.245
Intercept	2.908	2.799	2.689
Sample Size	92	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	- .219	- .675	.500
Curvelinear Trend	- .001	- .003	.997

TABLE 4.30. ANALYSIS OF COVARIANCE OF THE DPW RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	.485	.242	.485	.616
Zero Slope	2	5.740	2.870	5.739	.003
Error	164	82.011	.500		
Equality of Means (w/o Covariates)	2	.522	.261	.494	.610
Error	166	87.752	.528		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.007	.537	1.873	.062
Integration Level	.089	.269	.332	.739

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.946	3.979	4.078
Adj. Mean	3.922	4.052	4.075
(Std. Error)	.096	.137	.203
Intercept	- .113	.017	.039
Sample Size	92	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.153	.569	.569
Curvelinear Trend	- .108	- .300	.764

TABLE 4.31. ANALYSIS OF COVARIANCE OF THE DRP RATINGS STRATIFIED BY EXPOSURE LEVEL (CONTROL GROUPS OMITTED)

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	4.639	2.319	2.410	.092
Zero Slope	2	4.420	2.210	2.296	.103
Error	164	157.81	.962		
Equality of Means (w/o Covariates)	2	4.703	2.351	2.406	.093
Error	166	162.23	.977		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	1.575	.745	2.112	.036
Integration Level	-.526	.373	-1.409	.160

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	4.184	3.750	4.011
Adj. Mean	4.035	3.671	4.383
(Std. Error)	.134	.190	.281
Intercept	-1.397	-1.761	-1.049
Sample Size	92	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	.348	.934	.351
Curvelinear Trend	1.075	2.150	.033

TABLE 4.32. ANALYSIS OF COVARIANCE OF THE MEAN OF ALL DIFFICULTY RATINGS

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	.189	.094	.227	.797
Zero Slope	2	4.516	2.258	5.409	.005
Error	164	68.456	.417		
Equality of Means (w/o Covariates)	2	.814	.407	.926	.398
Error	166	72.972	.439		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Class Type	.839	.490	1.710	.089
Integration Level	.110	.245	.451	.652

TABLE OF MEANS

Exposure Level	(3)	(4)	(5)
Mean	3.933	3.757	3.932
Adj. Mean	3.918	3.829	3.909
(Std. Error)	.088	.125	.185
Intercept	.505	.416	.496
Sample Size	92	33	44

COMPARISON TESTED	VALUE OBSERVED	T-STAT.	SIGNIF.
Linear Trend	- .009	- .038	.969
Curvelinear Trend	.169	.514	.607

questionnaires at every site. For this reason the questionnaire data was pooled before analysis. The specific procedure used and the rationale on which it is based are presented under the analysis heading.

The overall data structure appears in figure 4.11. For purposes of problem comparison, only changes in the general interest rating or difficulty ratings were considered relevant. Consequently, the I and D sets for this factor each consist of one score per person: the mean rating for the set concerned.

Hypotheses. There were no a priori hypotheses for this dimension. Rather, we planned a series of post hoc comparisons to determine whether there were significant differences in overall effectiveness among problems. Overall effectiveness in this context was defined as the composite score for each variable set. As before, the composite score for the proposal and final report sets were the sum of the variable scores in the set, but for interest and difficulty the composites were the mean of the variables in the set.

Analyses. Ideally each of the four composite scores would have been analyzed using analysis of covariance to remove the effects of exposure level, class type, and degree of integration. However, this proved to be impossible because the control variables were invariant for subjects using problem seven. Consequently, the following procedure was used. First a one-way analysis of variance was run, with subsequent comparisons among all pairs of means. Standard t tests were used in these comparisons, because type two errors were of more concern than type one. This procedure was considered statistically conservative.

Following each analysis of variance an analysis of covariance was run, with problem seven omitted, to determine whether the covariates changed the observed relationships. Conclusions were then based on inferences from both analyses.

Results for Proposals. The results of the analysis of variance of the composite proposal ratings appear in table 4.33. It was observed that problems six and seven were significantly lower than all

SAMPLING PROCEDURE	PROBLEM CONTENT	CONTROL VARIABLES	PROPOSAL SCORES	FINAL REPORT SUM	MEAN INTEREST RATINGS	MEAN DIFFICULTY RATINGS
IG	1	C_1	P_1	F_1	I_1	D_1
IG	2	C_2	P_2	F_2	I_2	D_2
IG	3	C_3	P_3	F_3	I_3	D_3
IG	4	C_4	P_4	F_4	I_4	D_4
IG	5	No replication: only a single project (team) available.				
IG	6	C_6 invariant	P_6	F_5	I_6	D_6
IG	7	C_7	P_7	F_6	missing	missing
IG	8	C_8	P_8	F_7	I_8	D_8

Key to Symbols Used Above

Symbol	Description
IG	Indicates that sampling was by intact groups.
C	The set of control variables for a group. There are two scores for each individual: the mean of his exposure level and the mean of his type of class.
P	The sum of ratings on commonly assigned project tasks (IP + CF + M + GE) for a group.
F	The sum of ratings on commonly assigned final report tasks (IP + CF + M + GE) for the group.
I	The set of interest ratings from the ORS questionnaire (ICR, IEV, IPW, IRP) for the group.
D	The set of difficulty ratings from the ORS questionnaire (DCR, DEV, DPW, DRP) for the group.

Figure 4.11. Diagramatic Summary of the Data Matrix Stratified by Problem Content. Only Subjects Who Had FEHR Training are Included.

TABLE 4.33. ANALYSIS OF VARIANCE OF THE COMPOSITE PROPOSAL RATINGS STRATIFIED BY PROBLEM CONTENT

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	6	9181.8	1530.3	5.606	.000
Within	207	56497.	272.93		
Total	213	65679.			

PROBLEM	N	MEAN	VARIANCE	STD. DEV.
1	30	61.067	133.93	11.573
2	112	59.455	373.13	19.317
3	24	61.000	101.48	10.074
4	26	64.846	86.375	9.293
6	12	47.167	203.24	14.256
7	5	24.000	1080.0	32.863
8	5	66.000	36.500	6.041

PROBLEMS COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	1.611	.474	.635
1 vs. 3	.066	.014	.988
1 vs. 4	- 3.779	- .853	.394
1 vs. 6	13.900	2.463	.014*
1 vs. 7	37.067	4.644	.000*
1 vs. 8	- 4.933	- .6182	.537
2 vs. 3	- 1.544	- .415	.678
2 vs. 4	- 5.390	-1.498	.135
2 vs. 6	12.289	2.448	.015*
2 vs. 7	35.455	4.695	.000*
2 vs. 8	- 6.544	- .866	.387
3 vs. 4	- 3.846	- .822	.411
3 vs. 6	13.833	2.368	.018*
3 vs. 7	37.000	4.555	.000*
3 vs. 8	- 5.000	- .615	.538
4 vs. 6	17.679	3.066	.002*
4 vs. 7	40.846	5.063	.000*
4 vs. 8	-1.153	- .143	.886
6 vs. 7	23.167	2.634	.009*
6 vs. 8	-18.833	-2.141	.033*
7 vs. 8	-42.000	-4.019	.000*

other problems, but there were no significant differences among the other problems.

The results of the analysis of covariance of the proposal ratings appear in table 4.34. It was observed that the means for problems three, four, six and eight were adjusted upwards considerably. As a result, the adjusted means for problems three, four, and eight were significantly higher than the other means. In addition, mean eight was now significantly greater than mean two, which was adjusted slightly downwards.

Results for Final Reports. The results of the analysis of variance of the composite final report ratings appear in table 4.35. It was observed that the means for problem seven was significantly lower than those for all other problems. Also, problems one, three, four, and eight were significantly higher than those of problems two, six, and seven. There were no significant differences within the high means.

The results of the analysis of covariance are summarized in table 4.36. It was observed that there was only one change in order: the adjusted mean for problem six was larger than the adjusted mean for problem two. However, the difference was not significant. The means for problems six and two were significantly lower than any of the others; and problem one was significantly lower than problems four and three. There were no significant differences among the remaining means.

Results for Interest. The analysis of variance of the composite interest ratings are summarized in table 4.37. It was observed that the overall F was not significant (p = .0644). Similarly, the overall F for the analysis of covariance summarized in table 4.38 was not significant (p = .0814). It was concluded that there were no significant differences among problem means for the interest ratings.

Results for Difficulty Ratings. The analysis of variance for difficulty ratings appear in table 4.39, and the subsequent analysis of covariance appears in table 4.40. Again, the overall F ratios for both analyses were not significant (p = .1127 and .3337 respectively).

TABLE 4.34. ANALYSIS OF COVARIANCE OF THE COMPOSITE PROPOSAL RATINGS STRATIFIED BY PROBLEM CONTENT AND COVARIED ON INTEGRATION, EXPOSURE, AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	5	7915.7	1583.1	12.757	.000
Zero Slope	3	27309.	9103.1	73.356	.000
Error	195	24199.	124.09		
Equality of Means (w/o Covariates)	5	3445.7	689.14	2.649	.024
Error	198	51508.	260.14		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Integration	19.818	2.101	9.432	.000
Exposure	.487	.122	3.992	.000
Class Type	.678	.175	3.860	.000

TABLE OF MEANS

Problem	(1)	(2)	(3)	(4)	(6)	(8)
Mean	61.067	59.455	61.000	64.846	40.857	66.000
Adj. Mean	59.611	55.388	69.940	72.045	50.486	72.007
(Std. Error)	2.159	1.160	2.392	2.293	4.289	5.008
Intercept	-37.863	-42.085	-27.533	-25.428	-46.988	-25.467
Sample Size	30	112	24	26	7	5

PROBLEMS COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	4.222	1.657	.099
1 vs. 3	-10.329	-3.268	.001*
1 vs. 4	-12.434	-4.005	.000*
1 vs. 6	9.124	1.916	.056
1 vs. 8	-12.396	-2.287	.023*
2 vs. 3	-14.552	-5.227	.000*
2 vs. 4	-16.657	-6.202	.000*
2 vs. 6	4.902	1.083	.280
2 vs. 8	-16.618	-3.205	.001
3 vs. 4	- 2.104	- .665	.506
3 vs. 6	19.454	4.064	.000*
3 vs. 8	- 2.066	- .376	.706
4 vs. 6	21.559	4.540	.000
4 vs. 8	.038	.007	.994
6 vs. 8	-21.521	-3.290	.001

TABLE 4.35. ANALYSIS OF VARIANCE OF THE COMPOSITE FINAL REPORT RATINGS STRATIFIED BY PROBLEM CONTENT

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	6	32360.	5393.4	18.826	.000
Within	142	40681.	286.49		
Total	148	73041.			

PROBLEM	N	MEAN	VARIANCE	STD. DEV.
1	30	86.533	205.57	14.338
2	48	62.792	310.64	17.625
3	24	92.958	181.09	13.457
4	26	93.154	236.94	15.393
6	12	65.000	285.82	16.906
7	5	46.200	1598.7	39.984
8	4	90.750	164.25	12.816

PROBLEMS COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	23.742	6.026	.000*
1 vs. 3	- 6.425	- 1.386	.167
1 vs. 4	- 6.620	- 1.459	.146
1 vs. 6	21.533	3.724	.000*
1 vs. 7	40.333	4.933	.000*
1 vs. 8	- 4.216	- .468	.640
2 vs. 3	-30.167	- 7.129	.000*
2 vs. 4	-30.362	- 7.366	.000*
2 vs. 6	- 2.208	- .404	.686
2 vs. 7	16.592	2.086	.038*
2 vs. 8	-27.958	- 3.174	.001*
3 vs. 4	- .195	- .040	.967
3 vs. 6	27.958	4.672	.000*
3 vs. 7	46.758	5.619	.000*
3 vs. 8	2.208	.241	.809
4 vs. 6	28.154	4.766	.000*
4 vs. 7	46.954	5.680	.000*
4 vs. 8	2.403	.264	.791
6 vs. 7	18.800	2.086	.038*
6 vs. 8	-25.750	- 2.635	.009*
7 vs. 8	-44.550	- 3.923	.000*

TABLE 4.36. ANALYSIS OF COVARIANCE OF COMPOSITE FINAL REPORT RATINGS STRATIFIED BY PROBLEM CONTENT AND COVARIED ON INTEGRATION, EXPOSURE, AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	5	21106.	4221.2	21.677	.000
Zero Slope	3	8277.2	2759.1	14.169	.000
Error	130	25315.	194.73		
Equality of Means (w/o Covariates)	5	27649.	5529.8	21.894	.000
Error	133	33592.	252.57		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Integration	20.206	3.828	5.277	.000
Exposure	.330	.219	1.505	.134
Class Type	.197	.290	.679	.498

TABLE OF MEANS

Problem	(1)	(2)	(3)	(4)	(6)	(8)
Mean	86.533	62.792	92.958	93.154	58.571	90.750
Adj. Mean	80.033	65.016	95.533	93.886	60.523	89.184
(Std. Error)	2.984	2.369	2.882	2.871	5.414	6.997
Intercept	14.850	- .167	30.350	28.702	- 4.660	24.001
Sample Size	30	48	24	26	7	4

PROBLEMS COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	15.017	3.582	.000*
1 vs. 3	-15.500	- 3.657	.000*
1 vs. 4	-13.853	- 3.403	.000*
1 vs. 6	19.511	3.193	.001*
1 vs. 8	- 9.151	- 1.196	.233
2 vs. 3	-30.517	- 8.169	.000*
2 vs. 4	-28.870	- 7.326	.000*
2 vs. 6	4.493	.735	.463
2 vs. 8	-24.168	- 3.291	.001*
3 vs. 4	1.647	.410	.682
3 vs. 6	35.011	5.757	.000*
3 vs. 8	6.348	.837	.403
4 vs. 6	33.363	5.601	.000*
4 vs. 8	4.701	.620	.536
6 vs. 8	-28.662	- 3.227	.001*

TABLE 4.37. ANALYSIS OF VARIANCE OF COMPOSITE INTEREST RATINGS STRATIFIED BY PROBLEM CONTENT

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	5	4.779	.955	2.146	.064
Within	122	54.334	.445		
Total	127	59.114			

PROBLEM	N	MEAN	VARIANCE	STD. DEV.
1	24	4.542	.442	.665
2	54	4.983	.344	.587
3	21	4.561	.470	.685
4	20	4.762	.687	.829
6	5	4.868	.730	.854
7	0			
8	4	4.584	.163	.404

TABLE 4.38. ANALYSIS OF COVARIANCE OF COMPOSITE INTEREST RATINGS STRATIFIED BY PROBLEM CONTENT AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	5	4.380	.876	2.014	.081
Zero Slope	2	2.147	1.073	2.469	.088
Error	120	52.187	.434		
Equality of Means (w/o Covariates)	5	4.779	.955	2.146	.064
Error	122	54.334	.445		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	- .126	.236	- .533	.595
Class Type	.433	.245	1.766	.079

TABLE OF MEANS

Problem	(1)	(2)	(3)	(4)	(6)	(8)
Mean	4.542	4.983	4.561	4.762	4.868	4.584
Adj. Mean	4.498	4.968	4.588	4.804	4.919	4.623
(Std. Error)	.136	.092	.148	.148	.295	.330
Intercept	3.154	3.623	3.243	3.459	3.574	3.278
Sample Size	24	54	21	20	5	4

TABLE 4.39. ANALYSIS OF VARIANCE OF THE COMPOSITE DIFFICULTY RATINGS STRATIFIED BY PROBLEM CONTENT

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Between	5	4.825	.965	1.825	.112
Within	122	64.488	.528		
Total	127	69.314			

PROBLEM	N	MEAN	VARIANCE	STD. DEV.
1	24	3.798	.431	.656
2	54	4.157	.542	.736
3	21	3.780	.695	.834
4	20	3.690	.537	.732
6	5	3.930	.422	.649
7	0			
8	4	4.009	.009	.096

TABLE 4.40. ANALYSIS OF COVARIANCE OF THE COMPOSITE DIFFICULTY RATINGS STRATIFIED BY PROBLEM CONTENT AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	5	2.962	.592	1.158	.333
Zero Slope	2	3.129	1.564	3.060	.050
Error	120	61.359	.511		
Equality of Means (w/o Covariates)	5	4.825	.965	1.825	.112
Error	122	64.488	.528		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.204	1.275	.160	.873
Class Type	.803	.326	2.459	.015

TABLE OF MEANS

Problem	(1)	(2)	(3)	(4)	(6)	(8)
Mean	3.798	4.157	3.780	3.690	3.930	4.009
Adj. Mean	3.745	4.120	3.844	3.753	4.006	4.080
(Std. Error)	.147	.114	.165	.167	.344	.370
Intercept	- .148	.226	- .049	- .140	.112	.186
Sample Size	24	54	21	20	5	4

It was concluded that there were no significant differences among problems means for the difficulty ratings.

Summary of Results for the Problem Factor. The pattern of results from all eight analyses just presented is displayed in table 4.41. There was persistent tendency for the subjects in problems two, six, and seven to obtain lower composite ratings on the proposal and final report variables. Subjects from problems three, four, and eight consistently obtained higher ratings when covaried exposure level, class type and integration level, and problem one subjects fell in between the other two. However, these results cannot be considered definitive because differences among problems were confounded with the uncontrolled effects of differential instruction and differences in initial ability. The last is particularly important for problems one and two which were extensively used with introductory classes containing a large proportion of statistical neophytes -- many with an anti-research bias. Further research is needed to determine whether any of the problems is intrinsically inferior or superior for this purpose.

The results reported in table 4.41 suggest no clear pattern with respect to the interest and difficulty ratings. In view of the small differences in absolute size and the non-significance of the statistical analyses it was concluded that there was no evidence of differential effects of problems on these variables.

Factor Three: Type of Class

Description. There were three general class types included in the field study. A brief summary of each type appears below. A more detailed description was given at the beginning of the chapter. Type of class was intended as only a nominal scale: the order of listing is entirely arbitrary. The types of class were:

(1) A general research course offered as a service to a variety of other program. Most of the home programs aim to produce educational practitioners rather than researchers/evaluators.

(2) A specific research course which concentrates on methods most useful to a particular content and contains subjects interested in becoming researchers in that content area.

TABLE 4.41. COMPARISON OF THE ORDERED PROBLEM MEANS FROM THE ORIGINAL ANOVA WITH THE ADJUSTED MEANS FROM THE ANCOVA FOR FOUR VARIABLES. MEANS UNDERSCORED BY THE SAME LINE ARE NOT SIGNIFICANTLY DIFFERENT.

Proposal Ratings

ANOVA Means $\bar{P}_7 < \bar{P}_6 < \bar{P}_2 < \underline{\bar{P}_3 < \bar{P}_1 < \bar{P}_4 < \bar{P}_8}$

Adjusted ANCOVA Means $\hat{P}_6 < \underline{\hat{P}_2 < \hat{P}_1} < \underline{\hat{P}_3 < \hat{P}_4 < \hat{P}_8}$

Final Report Ratings

ANOVA Means $\bar{F}_7 < \underline{\bar{F}_2 < \bar{F}_6} < \underline{\bar{F}_1 < \bar{F}_8 < \bar{F}_3 < \bar{F}_4}$

Adjusted ANCOVA Means $\underline{\hat{F}_6 < \hat{F}_2} < \underline{\hat{F}_1 < {}}\underline{\underline{\hat{F}_8} < \hat{F}_4 < \hat{F}_3}$

Interest Ratings

ANOVA Means $\underline{\bar{I}_1 < \bar{I}_3 < \bar{I}_8 < \bar{I}_4 < \bar{I}_6 < \bar{I}_2}$

Adjusted ANCOVA Means $\underline{\hat{I}_1 < \hat{I}_3 < \hat{I}_8 < \hat{I}_4 < \hat{I}_6 < \hat{I}_2}$

Difficulty Ratings

ANOVA Means $\underline{\bar{D}_4 < \bar{D}_3 < \bar{D}_1 < \bar{D}_6 < \bar{D}_8 < \bar{D}_2}$

Adjusted ANCOVA Means $\underline{\hat{D}_1 < \hat{D}_4 < \hat{D}_3 < \hat{D}_6 < \hat{D}_8 < \hat{D}_2}$

(3) A program evaluation workshop intended primarily for part-time graduate students who are practicing educators. The clients were largely research novices who wished to increase their research/evaluation skills for practical reasons.

Data Matrix. The data set to be analyzed across class types is summarized in figure 4.12. Here, our interest focused on the pattern of proposal scores, so the entire set of P variables was included. However, only overall trends in interest or difficulty level were considered interprable: consequently, only the composite ratings were included for these variables. As usual, the final reports were evaluated in terms of a single composite score.

Hypotheses. The type factor was considered purely descriptive in all but two respects. (1) If, in fact, there is a difference in the content studied by specific research as opposed to general research classes, one would expect better problem definitions (IP ratings) and better conceptual frameworks (CF ratings) from the specialist group. (2) Since the workshop group actually received less training (fewer contact hours) than either of the other two groups, product ratings which compared favorably with the other groups could be considered evidence that the workshop was achieving its objectives. In particular, we hypothesized that the product ratings for this group would be as high as those of the general research type at the end of their first semester of training.

Analyses. Because of the interest in possible pattern differences, a multivariate analysis of variance with subsequent marginal (variable-by-variable) contrasts was conducted on the proposal scores. If a differential pattern was discovered of the multivariate analysis, analyses of covariance of each separate variable were planned. If no pattern differences were observed, we planned to proceed directly to the analysis of covariance of the remaining composite ratings: F, I, and D.

Results for Proposal Ratings. The multivariate analysis of variance for the proposal scores is summarized in table 4.42. It was observed that both type 1 and type 2 means appeared to be somewhat

SAMPLING PROCEDURE	TYPE OF CLASS	CONTROL VARIABLES	PROPOSAL SCORES	FINAL REPORT SUM	INTEREST RATINGS	DIFFICULTY RATINGS
IG	1	C_1	P_1	F_1	I_1	D_1
IG	2	C_2	P_2	F_2	I_2	D_2
IG	3	C_3	P_3	F_3	I_3	D_3

Key to Symbols Used Above

Symbol	Description
IG	Indicates that sampling was by intact groups.
C	The set of control variables for a group. There are two scores for each individual: the mean of his exposure level and the mean of his integration classification.
P	The set of ratings on commonly assigned proposal tasks (IP, CF, M, GE) for the group.
F	The sum of ratings on commonly assigned final report tasks (IP + CF + M + RC + GE) for the group.
I	The set of interest ratings from the ORS questionnaire (ICR, IEV, IPW, IRP) for the group.
D	The set of difficulty ratings from the ORS questionnaire (DCR, DEV, DPW, DRP) for the group.

Figure 4.12. Diagramatic Summary of the Data Matrix Stratified by Type of Class. Only Subjects Who Had FEHR Training are Included.

TABLE 4.42. RESULTS OF THE MULTIVARIATE ANALYSIS OF VARIANCE OF THE OBSERVED PROPOSAL RATINGS STRATIFIED BY CLASS TYPE

Equality of Means: DF = 8,436.00 F = 6.104 SIG. = .000

Alt. Test of Equality of Means: Max. Root = .204 SIG. = .000

TABLE OF MEANS

Class Type	(1)	(2)	(3)
IP	11.985	11.750	9.583
CF	10.422	14.750	6.416
M	25.706	28.125	19.917
GE	11.368	17.500	6.083
Sample Size	204	8	12

	OBSERVED VALUE	CRITICAL VALUES SIG. = .05	SIG. = .01
Comparisons For The Entire Set With Unit Weights On Each Variable			
1 vs. 2	-12.645	21.812	25.170
1 vs. 3	17.480	17.977	20.744
2 vs. 3	30.125	27.623	31.876*
Class Type Comparison by IP			
1 vs. 2	.235	4.345	5.014
1 vs. 3	2.402	3.581	4.132
2 vs. 3	2.166	5.503	6.350
Class Type Comparison by CF			
1 vs. 2	- 4.328	7.717	8.904
1 vs. 3	4.004	6.360	7.339
2 vs. 3	8.333	9.772	11.277
Class Type Comparison by M			
1 vs. 2	- 2.419	9.795	11.303
1 vs. 3	5.789	8.072	9.315
2 vs. 3	8.208	12.404	14.314
Class Type Comparison by GE			
1 vs. 2	- 6.132	5.249	6.057
1 vs. 3	5.284	4.326	4.992
2 vs. 3	11.417	6.648	7.671*

higher overall than type 3 means. In the paired comparisons for the entire set of variables (with unit weight on each variable) the type 2 vs. type 3 comparison was significant ($p < .05$), but the type 1 vs. type 3 comparison was just short of significance. The paired comparisons by variable yielded significant differences in the same direction, but only for the GE variable.

The hypothesized superiority of type 2 over type 1 classes for the IP and CF ratings was not supported: the direction of the difference was reversed for the two variables, and neither difference was significant.

Results for the Final Report Ratings. The analysis of covariance of the composite ratings for final reports are summarized in table 4.43. It was observed that the overall F ratio failed to reach significance ($p = .0841$): there were no significant differences among adjusted composite ratings (F) for final reports. The covariates were observed to be highly correlated with the F ratings, and resulted in a significant downward adjustment in the type 2 mean. This fact suggested that had a covariance analysis of the (multivariate) proposal scores been possible, the significant difference between type 2 and 3 may have washed out. Consequently, it was concluded that these results neither support nor deny the hypothesized differences in proposal and final report ratings.

Results for Interest Ratings. The analysis of covariance of the interest ratings appear in table 4.44. It was observed that there were highly significant differences among class types for both unadjusted and adjusted means. In both cases the types ranked 1, 2, and 3 in increasing order. The adjusted mean for type 1 was significantly lower than either of the others, but types 2 and 3 were not significantly different.

Results for Difficulty Ratings. The analysis of covariance of the composite difficulty ratings appears in table 4.45. Again it was observed that both the unadjusted and adjusted means for class types were ordered 1, 2, 3. However, the only significant difference in this case was between the two extremes: one and three.

TABLE 4.43. ANALYSIS OF COVARIANCE OF THE COMPCSITE FINAL REPORT RATINGS STRATIFIED BY CLASS TYPE AND COVARIED ON EXPOSURE AND INTEGRATION

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	1733.3	866.65	2.521	.084
Zero Slope	2	11143.	5571.7	16.213	.000
Error	135	46394.	343.66		
Equality of Means (w/o Covariates)	2	3915.9	1957.9	4.662	.011
Error	135	57537.	419.98		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.908	.252	3.603	.000
Integration	16.464	5.949	2.767	.006

TABLE OF MEANS

Class Type	(1)	(2)	(3)
Mean	78.642	100.38	74.167
Adj. Mean	79.536	90.894	71.546
(Std. Error)	1.819	8.531	6.169
Intercept	-11.975	-.616	-19.965
Sample Size	120	8	12

CLASS TYPES COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	-11.358	-1.243	.215
1 vs. 3	7.990	1.184	.238
2 vs. 3	19.348	2.194	.029

TABLE 4.44. ANALYSIS OF COVARIANCE OF THE COMPOSITE INTEREST RATINGS STRATIFIED BY CLASS TYPE AND COVARIED ON EXPOSURE AND INTEGRATION

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	9.766	4.883	8.068	.000
Zero Slope	2	21.126	10.563	17.452	.000
Error	195	118.02	.605		
Equality of Means (w/o Covariates)	2	10.003	5.001	7.081	.001
Error	197	139.15	.706		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	1.374	.341	4.025	.000
Integration	- .130	.228	- .571	.568

TABLE OF MEANS

Class Type	(1)	(2)	(3)
Mean	4.286	4.794	5.397
Adj. Mean	4.256	5.171	5.774
(Std. Error)	.060	.355	.370
Intercept	-1.549	- .633	- .013
Sample Size	185	8	7

CLASS TYPES COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	- .915	-2.461	.014
1 vs. 3	-1.518	-3.930	.000
2 vs. 3	- .602	-1.496	.136

TABLE 4.45. ANALYSIS OF COVARIANCE OF THE COMPOSITE DIFFICULTY RATINGS STRATIFIED BY CLASS TYPE AND COVARIED ON EXPOSURE AND INTEGRATION

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	2	2.599	1.299	3.004	.051
Zero Slope	2	.951	.475	1.100	.334
Error	195	84.335	.432		
Equality of Means (w/o Covariates)	2	5.320	2.660	6.144	.002
Error	197	85.286	.432		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.298	1.197	.249	.803
Integration	.143	.162	.884	.377

TABLE OF MEANS

Class Type	(1)	(2)	(3)
Mean	3.810	4.210	4.606
Adj. Mean	3.819	4.106	4.502
(Std. Error)	.049	.261	.275
Intercept	2.486	2.773	3.169
Sample Size	185	8	7

CLASS TYPES COMPARED	DIFF.	T-STAT.	SIGNIF.
1 vs. 2	- .287	-1.064	.288
1 vs. 3	- .683	-2.403	.017
2 vs. 3	- .395	-1.162	.246

Summary of Results for Class Type. Although some differences among project ratings reached significance, the overall pattern of results did not either support or deny the hypothesized superiority of type 2 classes over type 1 for any of the proposal ratings or for the final report rating. The workshop ratings on these variables were consistently lower than those of the other two class types, and these differences were generally significant. These results tend to contradict the proposition that workshops are as effective as the other class types in teaching these research/evaluation skills. However, the evidence cannot be considered conclusive because of the fact that differences in type were confounded with differences in initial ability: these were only partially controlled by the covariance.

Although no differences in interest or difficulty ratings were hypothesized, the results indicate that type 1 classes generated less interest than either of the other two. Type 1 classes rated research activities significantly less difficult than type 3 classes: type 2 was between the other two but not significantly different from either of them.

Factor Four: Integration

Description. The integration factor was included in the study to obtain some estimate of the differences in overall effectiveness between a FEHR project which was integrated into the curriculum as opposed to a laboratory experience and one which was just an adjunct to classroom activities. There was a certain amount of integration in many of the evaluation classes. For example, the project was discussed extensively during class at both Michigan State University and Ohio State University. However, the most complete planned integration of project activities and classroom content occurred in the C655-C656 course sequence at the University of Michigan during the 1973-74 academic year. Since the task of ranking the other sessions on a degree-of-integration scale seemed almost insuperable, we decided to use a two category classification: (1) incomplete integration, and (2) complete integration. The 1973-74 class was classified in the second group, and all other classes in the first group.

Data Matrix. Since overall effects rather than differences in patterns were of interest in the integration factor, only the composite ratings were included in the data matrix to be analyzed. These data, stratified by integration level are summarized diagrammatically in figure 4.13.

Hypotheses. It was hypothesized that the incompletely integrated class would achieve lower proposal and final report ratings, and would be less interested in research than the completely integrated classes. The difficulty ratings were not included in this hypothesis, and were included for descriptive reasons only. Summarized in symbolic form, the hypotheses were:

(1) $\bar{P}_1 < \bar{P}_2$

(2) $\bar{F}_1 < \bar{F}_2$

(3) $\bar{I}_1 < \bar{I}_2$

(4) $\bar{D}_1$ (?) $\bar{D}_2$

Analyses. To control for the effects of the exposure and type of class factors, an analysis of covariance of each of the four dependent variables was conducted. As in previous analyses, the covariate scores for each individual consisted of the mean scpres on the variable being analyzed for that individual's exposure level and class type.

Results for Proposal Ratings. The results of the analysis of covariance of the composite proposal ratings appears in table 4.46. It was observed that, as hypothesized, the adjusted mean rating for the incompletely integrated classes was significantly smaller than the adjusted mean for the integrated classes.

Results for Final Report Ratings. The results of the analysis of covariance of the composite final report ratings appear in table 4.47. It was observed that, as before, the adjusted mean for the incompletely integrated classes was significantly smaller than the adjusted mean for the integrated classes.

Results for Interest Ratings. The summary of the analysis of covariance of the composite interest ratings appears in table 4.48.

SAMPLING PROCEDURE	INTEGRATION LEVEL	CONTROL VARIABLES	PROPOSAL SUMS	FINAL REPORT SUMS	INTEREST MEANS	DIFFICULTY MEANS
IG	1	C_1	P_1	F_1	I_1	D_1
IG	2	C_2	P_2	F_2	I_2	D_2

Key to Symbols Used Above

Symbol	Description
IG	Indicates that sampling was by intact groups.
C	The set of control variables for a group. There are two scores for each individual: the mean of his exposure level and the mean of his type of class.
P	The sum of ratings on commonly assigned proposal tasks (IP, CF, M, GE) for a group.
F	The sum of ratings on commonly assigned final report tasks (IP, CF, M, RC, GE) for the group.
I	The mean of interest ratings from the ORS questionnaire (ICR, IPE, IPW, IRP) for the group.
D	The mean of difficulty ratings from the ORS questionnaire (DCR, DPE, DPW, DRP) for the group.

Figure 4.13. Diagramatic Summary of the Data Matrix Stratified by Integration Level. Only Subjects Who Had FEHR Training are Included.

TABLE 4.46. ANALYSIS OF COVARIANCE OF THE COMPOSITE PROPOSAL RATINGS STRATIFIED BY INTEGRATION LEVEL AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	1	6793.7	6793.7	42.368	.000
Zero Slope	2	6287.3	3143.6	19.605	.000
Error	201	32230.	160.35		
Equality of Means (w/o Covariates)	1	16451.	16451.	86.703	.000
Error	203	38518.	189.74		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.492	.137	3.593	.000
Class Type	.900	.182	4.947	.000

TABLE OF MEANS

Integration Level	(1)	(2)
Mean	52.333	70.432
Adj. Mean	54.236	67.902
(Std. Error)	1.262	1.489
Intercept	-28.354	-14.688
Sample Size	117	88

TABLE 4.47. ANALYSIS OF COVARIANCE OF THE COMPOSITE FINAL REPORT RATINGS STRATIFIED BY INTEGRATION LEVEL AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	1	2651.2	2651.2	7.726	.006
Zero Slope	2	6903.1	3451.6	10.059	.000
Error	136	46664.	343.12		
Equality of Means (w/o Covariates)	1	7885.9	7885.9	20.316	.000
Error	138	53567.	388.17		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.997	.231	4.313	.000
Class Type	.681	.329	2.064	.040

TABLE OF MEANS

Integration Level	(1)	(2)
Mean	76.086	96.000
Adj. Mean	77.210	90.568
(Std. Error)	1.769	4.278
Intercept	-55.403	-42.045
Sample Size	116	24

TABLE 4.48. ANALYSIS OF COVARIANCE OF THE OVERALL MEANS OF INTEREST RATINGS STRATIFIED BY INTEGRATION LEVEL AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	1	10.558	10.558	16.317	.000
Zero Slope	2	3.478	1.739	2.688	.070
Error	196	126.81	.647		
Equality of Means (w/o Covariates)	1	18.858	18.858	28.658	.000
Error	198	130.29	.658		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	-2.213	1.445	-1.531	.127
Class Type	.415	.410	1.013	.312

TABLE OF MEANS

Integration Level	(1)	(2)
Mean	4.156	4.844
Adj. Mean	4.130	4.912
(Std. Error)	.077	.151
Intercept	11.065	11.846
Sample Size	145	55

TABLE 4.49. ANALYSIS OF COVARIANCE OF THE OVERALL MEANS OF DIFFICULTY RATINGS STRATIFIED BY INTEGRATION LEVEL AND COVARIED ON EXPOSURE AND CLASS TYPE

Source	DF	SUM OF SQUARES	MEAN SQUARE	F-STAT.	SIGNIF.
Equality of Adj. Cell Means	1	.315	.315	.733	.392
Zero Slope	2	2.794	1.397	3.245	.041
Error	196	84.357	.430		
Equality of Means (w/o Covariates)	1	3.455	3.455	7.850	.005
Error	198	87.151	.440		

TABLE OF COEFFICIENTS

Covariate	COEFF.	STD. ERROR	T-STAT.	SIGNIF.
Exposure	.342	1.178	.290	.771
Class Type	.818	.334	2.446	.015

TABLE OF MEANS

Integration Level	(1)	(2)
Mean	3.773	4.068
Adj. Mean	3.817	3.952
(Std. Error)	.063	.123
Intercept	- .647	- .512
Sample Size	145	55

It was observed that both the unadjusted and adjusted means differed significantly in the direction hypothesized ($p < .00005$, and $p = .0001$ respectively).

Results for Difficulty Ratings. The summary of the analysis of covariance of the composite difficulty ratings appears in table 4.49. It was observed that the unadjusted means differed significantly in favor of integration. However, the difference in adjusted means, although still in the same direction, was much smaller, and non-significant ($p = .3927$).

Summary of Results for Integration. The overall results support the hypothesis that the quality of proposals and final reports and the overall interest in research/evaluation activities will improve when the FEHR research project is integrated into the curriculum rather than being a separate and adjunct experience. However, integration was found to have no significant effect on difficulty ratings.

CHAPTER 5

SUMMARY AND CONCLUSIONS

The purpose of this chapter is to summarize and synthesize the material presented in previous chapters in order to answer four major questions: (1) Do the FEHR materials satisfy the terms of the contract?, (2) Did the FEHR system accomplish its educational objectives?, (3) Are the eight problems equally effective? and (4) What are the implications of these results for the dissemination and use of the FEHR system? On the following pages, the evidence respecting each of these questions is presented, under the appropriate heading, in the order listed.

Satisfaction of the Contract

It was pointed out in Chapter 2 that the present FEHR-PRACTICUM materials differ in six ways from the specifications contained in the contract addendum dated September 28, 1971.

1. The number of In Service Training Units actually produced was five instead of the four itemized in the contract.
2. The programmed materials in the *Players' Instructions* booklet were substituted for proposed Players' Oreintation Booklet. This eliminated the necessity for a training session prior to solving an actual problem.
3. The role of the message generator was de-emphasized. Rather than a necessary game component it is now an option to be used at the discretion of the Game Manager. Instructions and suggested messages appear only in the Game Manager's Manual. This action was deemed necessary because of the information overload experienced by many players in the earlier version.
4. The number of problems was reduced from ten to eight. One reason for this was that problem development and evaluation turned out to be a much more formidable task than had been anticipated. A second reason was

the extreme difficulty experienced by our staff in identifying a workable problem in the areas which had been proposed -- four separate areas were investigated in the course of the project. But the third and most important reason was the potentially greater return for concentrating our efforts on methods of producing multiple versions of each problem (see point 5, below).

5. The number of different versions (i.e., different difficulty levels) of each problem was increased from the proposed four to an almost infinite number by allowing each game manager/instructor to define his own assigned problem via a "checklist of assigned tasks". The total number of fundamentally different problems available with the present system far exceeds the 40 which would have resulted from the original proposal. More important, the increased flexibility increases the number of potential clients for the system.
6. The scope of the Game Manager's Manual was greatly increased. It is obvious from the above description that the present FEHR materials exceed the contract specifications on every important dimension. We consequently conclude that the requirements of the contract with respect to materials are fully satisfied.

Accomplishment of Objectives.

The purpose of this section is to summarize, integrate, and interpret the evaluation results with respect to the specific goals of the FEHR-PRACTICUM system. It is organized in eight sections corresponding to the eight educational objectives listed at the beginning of Chapter 4. The first three objectives are concerned with achievement, objectives four to seven are concerned with the overall utility of the system, and objective eight concernsthe overall adaptability of the system.

The discussion below is organized by objectives with an extended discussion of results for the achievement and interest goals. Within objectives the presentation will consist of a statement of the objective, an identification of the appropriate dependent variables, and a summarization and interpretation of the results for each of three critical comparisons. These were concerned with the effects of: (1) varying the exposure to the FEHR system from zero (the control condition) to an extended exposure to all aspects of the system (the extensive-intensive condition), (2) the type or purpose of the course or session (required research competency, research specialist, program evaluation workshop), and (3) integrating FEHR with an existing curriculum (integrated vs. non-integrated practicums). For each contrast, a summary of the evidence is presented, followed by a brief discussion and interpretation. The original design included comparisons among the eight problems on an objective-by-objective basis. However, this proved to be impractical because of missing data and unbalanced replication. Consequently the relative effectiveness of the various problems is assessed in a separate section.

Objective 1.

To improve achievement in the content areas usually associated with research/evaluation training.

The evidence respecting this objective comes from both the controlled experimental study and the field trials. Because these involve different dependent variables, the findings are presented separately.

Experimental Study Results. In the experimental study, overall achievement was assessed by the final test scores (FT). Students' perception of their learning was assessed by their mean score on the competency items (the MC score) from the Self Assessment of

Evaluation and Research Skills questionnaire. The SARES questionnaire was administered to the experimental subjects again at the end of the second semester to provide an estimate of the carryover effects of an early FEHR experience. Finally, the methods score (M) from the Project Rating Scale for both proposals and final reports was used as a measure of the practical application of research/evaluation knowledge and skill. Since the control group did not complete a project in the first semester M scores were available only for the experimental group in semester one. In semester two - M scores were available for both the experimental and control groups and for a small group of new students as well.

Amount of Exposure. The test for a linear increase in the dependent variables with increasing levels of exposure to FEHR was highly significant for both the FT&MC scores ($p = .02$ in both cases). The linear trend for the M scores was also highly significant ($p<.01$) both for the original experimental subjects and when new students were pooled with experimental subjects by degree of exposure.

Classtype and Integration. Tests of the effects of class type and integration were not possible in the experimental study.

Field Trial Results. In the field trials, a direct measure of performance in research design and data analysis was again given by the M subscore on the product rating scale for proposals and final reports. An indirect measure of performance in research evaluation content during the field trials was provided by student ratings of goals 4, 5, and 6 on the goal assessment questionnaire (GAQ). These goals dealt with the integration of measurement-research-data analysis knowledge and skill, the computation and use of specific statistics (e.g., t tests), and the ability to use appropriate computer programs respectively. The score analyzed represented the students' perception of the degree to which he or she had achieved these three goals. Unfortunately, the GAQ was available only for courses at the University of Michigan.

Amount of Exposure. The results of the test for linear trend were entirely consistent with those from the experimental study. Within the multivariate analysis of variance (Table 4.12, p. 178) the linear trend for M was highly significant. An impressive, but somewhat misleading result was the fact that participants' M scores were on the average, about double those of a control group consisting of nine typical PhD proposals. However, this discrepancy cannot be attributed entirely to differences in competence because rigorous criteria are frequently not applied at the proposal stage. Nevertheless, it is heartening to know that the performance of the FEHR participants compared very favorably with that of a group which can be assumed to be both reasonably competent and well motivated.

In the case of the goal achievement scores there was not a pure linear relationship with exposure level: the lowest and middle levels obtained similar mean ratings and the high exposure group a somewhat better rating. However, the multivariate test for linearity was not significant either for combined variables or variable by variable. To check for the possible confounding of results by the effects of integration level and class type (purpose) a subsequent analysis of covariance of the overall mean goal ratings was conducted using factors. The results of this analyses produced means of 3.55, 3.37 and 3.78 for the three exposure levels: the test for linear trend yielded a probability of .055, and the test for curvilinear trend a probability of .002. These combined results were considered inconclusive with respect to perceived goal attainment per se but not inconsistent with the hypothesized monotonic increase in achievement.

Class Type. Although the variable-by-variable comparisons within the MANOVA did not reach significance, the differences among the M means for the three class types were considerable: 25.7 for the courses required as part of a general PhD research competency, 28.1 for the research specialists, and 19.9 for the program evaluation workshop. Since the score pattern was consistent and the overall differences were significant subsequent univariate analyses of composite scores were run with exposure and integration levels controlled by covariance. These yielded highly significant differences among all three groups. From lowest to highest M scores the groups ranked: program evaluation, general research competency, and research specialists. This, of course, was to be expected: classes emphasizing formal research techniques ought to score better on variables measuring achievement in that area. What was most gratifying was the absolute size of the proposal ratings for the in-service workshop type of classes. With no formal instruction and generally, no prior research training, these people were able to write proposals and reports which scored about two-thirds as high as the formally trained groups. Since the class type variable involved off-campus courses, no GAQ scores were available for this comparison.

Integration. As hypothesized, the composite scores for proposals and final reports (predominantly M scores) yielded significantly higher ($p<.005$) mean scores for integrated courses. Classes in which the FEHR project was fully integrated into the classroom content achieved a mean score of 70.4 (67.9 after covariance adjustment for differences in exposure level and class type) while classes using FEHR as a free-standing practical experience achieved a mean of 52.3 (54.2 after adjustment). This was the largest absolute difference obtained in the analysis of product ratings.

Conclusions. The overall pattern of evidence suggests that exposure to the FEHR system was remarkably successful in improving achievement in subject matter related to traditional content of courses in research design and data analysis. Throughout both the experimental and field studies greater exposure to FEHR resulted in greater achievement. Furthermore, this result held whether the purpose of the session was specialized research training, a required research course, or an in-service workshop in program evaluation techniques. The second general conclusion was that the FEHR experience was far more effective when it was an integral part of a course or program of courses in which research methods were taught on a formal basis than when it was used as an independent practical-application experience.

The differing achievement of various class types was interpreted as evidence that shifts in emphasis did occur in these studies: classes placing more emphasis on formal research techniques scored better on variables measuring achievement in that area. Consequently, this factor was controlled when making other contrasts. A more important result of the class type comparison was the suggestion that formal research techniques were learned surprisingly well by subjects in the informal workshop sessions. Although the evidence in this study was not extensive enough to permit firm conclusions, it appears that the FEHR experience motivates and facilitates the learning of formal research techniques through self study.

Objective 2.

To develop the ability to write proposals and final reports which are explicit, operational, and sufficiently comprehensive to permit replication.

The most compelling evidence of the quality of the proposals and final reports produced by FEHR trainees appears in Chapter 3. The illustrative report for problem 2 (REMAR) is an exact copy of an actual final report submitted by a member of a class which had level 5 (the highest) exposure to FEHR. The dependent variables relevant to this objective are the various subscores on the product rating sheet (PRS). The sample report is about typical of the mean quality of the work in that level as evidenced by its proposal ratings (IP=11/20; CF=12/24; M=31/40; and GE=13/20) compared to the mean ratings for level 5 given in table 4.12 (13.226; 14.962; 30.755; and 13.962 respectively). Although the mean scores represent only about half the possible score, they are considerably better than the means for the control group of PhD. doctoral proposals (14.667; 11.778; 14.778; and 10.667 respectively), particularly for the Methods scores.

Amount of Exposure. The interest here is in the effects of increasing levels of exposure on the pattern of subscores (as opposed to the M score only in Objective 1).

The multivariate analysis of variance of the proposal ratings yielded a significant overall linear trend: PhD. proposals ranked lowest, then FEHR exposure levels 3, 4, and 5 in the order mentioned. On a variable by variable basis, however, it was obvious that the linearity held only within the FEHR treatments. The PhD. proposals were about equal to the better FEHR proposals for all but the method variable. Analyses of covariance of the composite proposal ratings and composite final report ratings yielded similar results.

Type of Class. The results for the type of class comparison yielded no significant differences in the pattern of product rating scores between required research classes and classes for research specialists. However, the scores for the workshop classes were significantly smaller than those for the other two types.

Integration. As reported in Objective 1, above, the mean composite product rating for the classes with an integrated FEHR project was significantly greater than that for those with non-integrated projects.

Conclusion. Collectively, the evidence strongly supports the conclusion that FEHR experience improves the trainee's ability to write proposals and final reports. Further, the evidence suggests that the more FEHR experience, the better -- at least within the range of time (up to 16 months) and number of separate projects (the maximum in this study was three projects) used in this study. On the basis of these data we recommend two problem experiences spread over two semesters (or the equivalent) for the usual PhD. research training sequence. However, additional research is needed to determine the optimum amount of experience to be provided.

Further indirect evidence of proposal quality is provided by the observation that the proportion of trainees who obtained a perfect score (5) for the "study is replicable" item on the product rating sheet was approximately 57%, 73%, and 90%, respectively, for the three FEHR exposure levels, as compared to 50% for the controls. Although the control comparison is dubious at best -- prevailing practice emphasizes the replicability criterion for completed dissertations but not for PhD proposals -- the absolute value of the statistic for the high exposure groups is impressive.

Additional support for the effectiveness of FEHR projects for developing proposal/report writing skills is provided by the ratings and written comments on the goal attainment questionnaire. These indicated that more than 90% of the trainees felt they had learned a great deal about proposal writing. But perhaps even more telling was the fact that better than three quarters of all respondents also indicated that they wished to learn more about the topic.

In summary, there is strong evidence that FEHR is a remarkably effective vehicle for developing proposal and report writing skills.

Objective 3.

To encourage field studies which feature control groups, multiple dependent variables, and an assessment of cost effectiveness.

We had initially intended to use the product rating scale items corresponding to control, multiple dependent variables, and cost effectiveness to assess achievement of this objective. Indeed, the differences among groups on these variables was considerable. However, examination of the assigned tasks revealed that univariate assessments were almost invariably performed because the class concerned had been restricted to one dependent variable, otherwise multivariate designs were used. Similarly, projects appeared to involve cost-effect assessments whenever this was assigned and virtually all projects involved a control group if one was possible in the problem concerned. Apparently, the mere presence of a control constituted an "assignment".

The effect of the highly visible "presence" of a variable in the simulated situation was particularly striking in the case of costs which are always printed out by the FEHR program. Even when classes were explicitly instructed to ignore experimental costs, they seemed unable to do so. Again and again the authors observed subjects in these classes arguing against certain test selections and experimental designs because they were "too expensive to be practical". Although not as prevalent, similar arguments were heard with respect to the inadvisability of making a decision on the basis of a single variable. In classes restricted to a single dependent variable, many teams requested that a different dependent variable be assigned to each team member. The rationale for this was usually based on the notion that gains on one variable could be offset by losses on another. For example, in the remedial arithmetic problem several teams felt that losses (i.e., arrested growth) in problem solving ability attributable to removal from class was just as important as gains in computational skill resulting from the remedial treatment. They were therefore unwilling to make a decision on the basis of computation alone even though the added variable meant additional work and the course requirements could have been entirely satisfied with a univariate assessment.

Unfortunately, there was no hard evidence for the attainment of objective three which could be attributed to the FEHR experience as opposed to the requirements imposed by the instructor and/or the implicit requirements of the problem description. A formal assessment of objective three would require a follow-up study to compare the research studies (especially the dissertations) of students who experienced FEHR with those who did not. Unfortunately, the time constraints of this project did not permit such comparisons to be made -- only a few students have reached the dissertation stage. Some additional support for this statement was provided by a post hoc examination of the mean responses to the control group question on the final test used in the experimental evaluation. The mean for this item increased uniformly with the amount of FEHR experience, but the difference was not significant ($p>.10$).

Conclusions. The evidence for the achievement of objective three was inconclusive. In the absence of a definitive follow-up study, the best that can be said is that the FEHR experience appears to alert subjects to the desirability of including control groups, multiple dependent variables, and cost effectiveness assessments in their program evaluation studies.

Objective 4.

To increase the interest in research and research methods.

Evidence for the attainment of this objective comes from three sources: (1) the attainment ratings given to this goal on the GAQ, (2) the mean interest rating from the SARES instrument (experimental study only), and (3) the mean of interest scores on the ORS questionnaire (field study only).

Amount of Exposure. The scores on the first three instruments are curvilinearly related to FEHR exposure. People with zero exposure (the control groups) tended to exhibit moderately high interest. This dropped significantly with the first exposure to research (either real or FEHR), then gradually climbed back up to exceed the control group at the extensive-intensive exposure level (level 5). In all cases, the relationship was significantly linear within the FEHR exposure levels.

The activity level, as measured by the number of reported voluntary pursuits of research-related tasks was not available for the control (zero experience) group. However, within the four FEHR exposure groups there was a linear increase in the proportion of people engaging in such activities: the observed percentages were 23%, 38%, 45%, and 48% for groups 2 to 5 respectively. Because of missing data and non-comparable formats, it was not feasible to test the significance of this trend, nor of individual differences between pairs. However, the data does lend increased credance to the rating-scale results.

The relatively high rating given to research tasks by neophytes was at first somewhat puzzling. However, after discussing the phenomenon with a number of trainees, a possible explanation occurred to us. In this age it is popular to revere and romanticize science -- hence an unrealistically high rating by the totally inexperienced. With the first experience comes the realization that research involves a great deal of hard work and an exacting routine. Disillusionment sets in and the ratings dip sharply. From this point on, the ratings are based on direct experience: the trend within FEHR exposure levels should therefore represent real changes in attitude.

This explanation is so consistent with common experience that we find it appealing. However, its acceptance is not critical to our case. Regardless of the persuasiveness of the above explanation, the pervasive linear relationship between various interest variables and degree of FEHR exposure within the experimental groups is sufficient to warrant a positive conclusion -- particularly in view of the fact that the non-FEHR research experience (group two in the field study) yielded the lowest interest means of all (see table 4.21).

Type of Class. The type of class comparison yielded insignificant differences for all except the ORS interest items (I). The means for the required research classes was significantly lower than the means for either the research specialists or the workshop classes, but there was no significant difference between specialists and workshop classes. However, within each class the linear relationship to FEHR exposure was maintained. It was concluded that the lower interest in class one was a function of its required status, and in no way related to the FEHR system per se.

Integration. None of the interest variables yielded a significant difference between classes with projects integrated into course content and those with independent projects.

Conclusions. Other things being equal, exposure to FEHR does indeed increase interest in research and research method. Further, within the range of exposures used in this study, the greater the exposure, the greater the interest produced. This relationship held equally well for required, specialist, and workshop classes.

The hypothesis that integration of the FEHR project into course content would increase interest was not supported: an independent project appears to be equally effective in stimulating interest.

Objective 5.

To increase the perceived relevance of the methods and practice of research/evaluation to (the trainees) educational role.

The evidence for attainment of this objective comes from the relevance scores on the SARES questionnaire which was administered in the experimental group only.

Amount of Exposure. The pattern of results almost exactly parallels those for the interest variables (see Tables 4.6, p. 164). The controls, who were not exposed to any practical experience in research perceived research methods as moderately relevant to their goals. With a minimum exposure to a practicum (only the routine items) this dropped somewhat, but the mean value remained moderately high. After the first experience, the perceived relevance of research/evaluation increased linearly with exposure to FEHR.

Type of Class and Integration Level. Neither of these comparisons were available in the experimental study.

Conclusion. Exposure to FEHR increases trainees' perception of the relevance of research/evaluation to their work role.

Objective 6 and 7. To foster positive attitudes toward the computer and team work.

The evidence for attainment of these objectives comes solely from written responses solicited from students. Consequently, none of the three contrasts was deemed appropriate. Only 200 students -- all at the University of Michigan -- were accessible to be polled; answers were received from 163 of them. Of these, 85% commented that the practicum experience had improved their regard for the computer and lessened their fear. Only 10% said it had not affected their view of the computer. The remainder had no opinion.

The team work question was answered by 139 respondents. Of these, 70% considered the team experience valuable and rewarding but fully 20% found it irksome; the remainder had no opinion.

Conclusion. The FEHR experience as given at the University of Michigan seems to be quite successful in fostering positive attitudes towards the computer. It is also successful in fostering a positive attitude towards group work in most people. However, a considerable minority were negative to the group experience. New methods of grouping should be tried with these people.

The seven specific objectives above were all assessed in terms of their degree of exposure of FEHR, with zero exposure used as a control group. It was also desirable to obtain assessments of two additional critical comparisons: problem content areas, and degree of curricular integration.

Problem Content.

Analyses of differences across the eight problems were hampered by statistical problems caused by missing data and radical heterogeneity of variance. Since problem five had only one complete formal usage of the revised problem, it was omitted from analysis. Because of missing data only overall analyses were attempted: it was not possible to assess the interactions between problem content and exposure level. However, covariance introduced another difficulty -- problem seven had to be dropped because of invariant covariate scores.

The analysis of covariance for the interest and difficulty scores yielded no significant differences among problems 1, 2, 3, 4, 6, and 8.

An examination of the mean proposal and final report ratings shows that problem seven was very low on both. This was considered natural, since things like control groups and valid treatment comparisons, which are heavily weighted on the rating sheet, are not available in the RMA problem. A similar restriction on the comparability of ratings exists for problem 6 (TQUES). Finally, the fact that problem 2 (REMAR) was used almost exclusively with neophytes taking a required course raised a question about whether its significantly lower mean score was validly related to the problem content *per se*. (For a diagrammatic summary of these results see table 4.41.)

The net result of the above considerations is that no meaningful conclusions about the relative effectiveness of the eight problems can be drawn from the tests and ratings. However, a number of practical questions about the relative utility of the problems can be answered from the evaluative comments by trainees, game managers and course instructors. Since these are presented in considerable detail in Chapter III, they are not repeated here: the highlights are presented under conclusions, below:

Conclusion. On the basis of practical experience and the feedback received from various FEHR users, the following conclusions about the relative utility of the eight problems seem warranted. See Chapter III for further details.

(1) All eight problems are effective as a vehicle for acquainting trainees with the program evaluation process.

(2) Trainees are especially interested in problems concerned with their (approximate) area of interest.

(3) The REMAR problem is especially well suited to general courses because almost all education students are familiar with remedial arithmetic as a content area. All of the other problems require specialized knowledge which the average education student does not have, and frequently does not need to know. Consequently, we recommend that these problems be used only with groups of subjects for which the particular contents will be of value.

(4) The READING problem needs to have its internal parameters adjusted to reduce the proportions of students at criterion on the objective referenced tests. (This task is currently underway.)

(5) The TQUES problem is useful for studying questionnaire analysis and construct validity procedures. It is also a useful vehicle for studying the practical and theoretical implications of student evaluation of courses via questionnaire items. It is not particularly suited to program evaluation as such.

(6) The BUSING problem is probably no longer of direct relevance given the current national status of the busing issue. However, it may still be useful to study the process involved in evaluating the effects of policy changes of this sort.

Implications for Dissemination and Use

The findings reviewed above provides compelling evidence that, correctly used, the FEHR system can be enormously useful in teaching research/evaluation skills. FEHR-PRACTICUM in its present flexible form has proven quite effective for creative instructors who are willing to adapt their methods to the problem solving mode which is inherently most compatible with the FEHR system. It seems equally apparent that it may not prove useful to instructors who are unwilling or unable to adapt their methods. The implications for dissemination derive from the pedagogical philosophy built into the FEHR system.

Some Philosophical Considerations. FEHR was designed to be a flexible pedagogical tool adaptable to many instructional purposes. To accomplish this aim, the problems were described in rather global terms, leaving the operational specification of the problem to the users. Thus, if an instructor/Game Manager desired his/her trainees to practice problem definition skills he/she could require the teams themselves to operationalize the problem. If, on the other hand, the instructor/Game Manager wished to concentrate on research design and analysis skills, he/she might provide an operational definition of the problem and ask the teams to work within it. Additional adjustment to the scope of the players' task could be made by restricting the number of treatments to be assessed, the number of variables to be considered, and/or the number and type of research subjects to be used.

Despite the conscious emphasis on adapting to an instructor's purposes, it might be a mistake to assume that FEHR is completely non-didactic. Like most instructional products, the FEHR-PRACTICUM system is an implicit operational statement of the instructional philosophy of its authors. There is a pervasive bias which tends to nurture a particular view of the research process and to encourage the use of some instructional practices while discouraging others. We believe that the optimal results can be achieved only if FEHR is used in a manner consistent with its basic structure. Consequently, the remainder of this section is devoted to an explication of the more important beliefs and principles upon which FEHR-PRACTICUM is based.

a. We believe that the empirical evaluation of educational programs is inherently a multidimensional process requiring the interrelation and synthesis of frequently conflicting information from a variety of sources. In our view, a single measure can almost never provide adequate assessment of educational effectiveness _per se_. In addition, the practical realities dictate that many factors other than a program's effectiveness in meeting an objective be considered. For example, the cost of a program and its degree of support among teachers, parents, and students must be taken into account. To complicate the process still further, there is always a host of irrelevant variables to divert the researcher/evaluator's attention from the important issues. In an attempt to capture some of this multidimensionality, each FEHR problem contains a variety of variables (tests) in each of several domains (attitudes, achievement, etc.), and several subgroups of subjects.

b. We are firmly convinced of the validity of the notion that we best learn research skills by doing research. In the area of evaluation and decision-oriented research, we would put the case even more strongly. One can learn to handle ambiguity and complexity only by working with ambiguous problems in a complex environment. Each FEHR problem is designed to provide this kind of experience. The problem definition supplied in the RFP is purposely broad and somewhat ambiguous, and there are always several treatments, many dependent variables (variables which change as a result of a treatment), and many moderator variables (variables which do not themselves change as a result of a treatment, but which change the effect of the treatment on one or more dependent variables).

c. We recognize that for novice trainees it may be pedagogically desirable to begin on a simplified problem. However, we consciously opted _not_ to provide simple problem descriptions with only two or three treatments and a single dependent variable. However, the Game

Manager or the players themselves may delimit the problem to provide an equivalent simplifying effect. It is our belief that a problem which is consciously delimited in the presence of complexity provides a more valid view of research, and consequently develops skills which are more likely to generalize to field research than would result from presenting only the delimited problem without the surrounding details.

d. The above view of the research/evaluation process suggests that there is no universal research method which can be learned in a relatively simple context (e.g., a laboratory), and later applied directly to practical problems in a variety of settings. Rather there are a variety of methods and techniques which must be combined, adapted, and synthesized to meet the idiosyncracies of a given practical problem. Since these combinations and adaptations frequently result in methods which differ in substantive ways from the originals, we call the resultant strategy an idiosyncratic research method.

The FEHR system provides for training in the development of idiosyncratic research strategies in two ways. First, the eight problems require vastly different research approaches. Second, within each problem it is possible to define the research objectives in several different ways, with each definition requiring a different research approach.

e. The need for idiosyncratic methods demands that programs to train researchers/evaluators emphasize the process by which a research strategy is developed rather than the strategy *per se.* For this reason the entire FEHR system is designed to create the desire to know and to provide an opportunity to discover.

One can best illustrate the discovery approach by examining its alternative. It is possible to use FEHR didactically. For example, a particular research strategy could be taught by "solving" a FEHR problem in class, and then asking trainees to practice that solution method using a different sample of subjects (e.g., a different school). While this sort of practice is undoubtedly useful, we do not believe that it takes full advantage of the system's power. Nor does it facilitate learning how to adapt a theoretical method to a practical need.

A less didactic procedure which is more consistent with the training needs would proceed as follows: First, trainees are allowed to struggle with a problem until they develop a need for the method to be taught (but not long enough to become overly frustrated). Second, the research method is taught utilizing an example different from the problem with which trainees are working. Third, trainees adapt the method to their own problem needs. We are convinced that this "discovery" approach will result in a greater depth of understanding and longer retention than more didactic procedures.

f. The discovery approach outlined above requires that a great deal of individualized instruction be available during the practicum. The FEHR consultants are intended to provide this service. In our experience, intensive team-by-team consultation provides far greater increments in learning than a comparable effort expended in large-class session--even though the latter method covers (at least superficially) far more material. To supplement the consultants, some users may wish to make a variety of programmed materials on research methods available to the players. Several examples of suitable materials are listed in the appendices of the Player's Instructions.

In any case, we believe it is a serious mistake to use the FEHR system to supplement an existing research/evaluation course without adopting appropriate instructional techniques.

g. The foregoing emphasis on multidimensionality and complexity encourages teams to devise studies involving data sets which are considerably larger than those found in the usual laboratory exercise. The opportunity to develop skills in this area is a feature of FEHR which ought to be exploited whenever possible.

h. Despite FEHR's admitted bias towards large data sets, the sheer size of the research populations, the number of available variables, and the redundancy of information (e.g., some problems have seven or eight intelligence tests) encourages the use of sampling for both subjects and variables. In most settings we would urge the game manager to provide further motivation in this direction by placing reasonable limits on budgets, number of subjects and number of variables.

i. The budgeting aspects of FEHR are considered an important and integral part of the practicum. More than any other element, costs motivate the players to plan their activities. Budgeting financial resources generalizes to budgeting of time and (nonfinancial) resources. In fact, it has been our experience that the various costs attached to treatments cause trainees to change their behavior even when they have been told to ignore costs. For example, most trainees refused to use the Stanford-Binet IQ test when they noticed its price ($12.65 per student) even though they were not being charged for it.

In respect to costs, it is important for the user to realize that there is an intricate non-linear relationship among test cost, reliability and total experimental cost. The experimental costs can only be compared by holding statistical power constant. To get an intuitive feel for this relationship, assume a matched experimental design in which there is a perfect correlation between the true scores of the matched elements. In this case, all the experimental error is attributable to test unreliability. Thus, given test A with a reliability of .91 for \$3 and test B with a reliability of .84 for \$2. Using test B the error variance would be SQRT(1-rel) or 4/3 times the error using test A. To maintain statistical power equivalent to that obtained with test A, we must use $4^2/3^2$ or 1.77 times as many people in the experiment. Thus using test B we would actually spend \$2 X 1.78 = \$3.45 for each \$3 using test A.

j. Finally, we believe that the team approach provides an added dimension of great value to the FEHR-PRACTICUM experience. The value is of two sorts. First, our experience shows that there is a tremendous amount of intra team teaching and learning during a FEHR project. Second, evaluative research in the practical world tends to be a team project. Consequently, any group-process skills learned during the practicum will have a direct and positive carryover. We urge instructors/game managers to use teams wherever possible. Our experience shows that the team size should not be smaller than three nor greater than five. Larger teams tend to break into subunits with one set of trainees doing most of the work. Smaller teams tend to have less verbal interaction and hence less opportunity to learn.

Some Difficulties: Need for Further Development

The most definitive finding of the evaluation was that FEHR projects are most effective when they are an integral part of a training curriculum. In the previous section we have spelled out some of the principles by which a beneficial integration can be accomplished. However, it should be obvious that this is a difficult and demanding task. Unfortunately, many would-be users of the FEHR system have neither the time nor inclination to make the necessary adaptations. What is required to reach these potential clients is a didactic and comprehensive programmed curriculum which utilizes all of FEHR's unique capacities to teach research/evaluation techniques and principles in a problem solving discovery made. The authors of these materials are currently exploring various methods of supporting this additional work.

REFERENCES

BUROS, Oscar K. The Sixth Mental Measurements Yearbook, 1965.

RESTA, Paul E. and BAKER, Robert L. Components of the Educational Research Proposal. New York: American Book Company, Van Nostrand Reinhold Company, 1972.

STAKE, Robert E. "Toward a Technology for the Evaluation of Educational Programs," AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally & Company, 1967.

STUFFLEBEAM, D.I., Foley, W.J., Gephart, W.J. Guba, E.G., Hammond, R.I., Merriman, H.O., and Provus, M.O. Educational Evaluation and Decision Making. Itasca, Illinois: F.E. Peacock Publishers, Inc., 1971.

TUCKMAN, Bruce W. Conducting Educational Research. New York: Harcourt Brace Jovanovich, Inc., 1972.

TYLER, Ralph W. "Changing Concepts of Educational Evaluation," AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally & Company, 1967.

APPENDIX 3A

BUDGET

Treatments

AUTOMATH

Set-up Cost/40 Subjects		$0
Maintenance Cost/40 Subjects/15 Weeks		12,000.00
Measurement - SATCOMP/40 Subjects		12.00
ITBSCONC/40 Subjects		8.00
	Sub-Total	$12,020.00

IRA

Set-up Cost/40 Subjects		$ 800.00
Maintenance Cost/40 Subjects/15 Weeks		6,000.00
Measurement - SATCOMP/40 Subjects		12.00
ITBSCONC/40 Subjects		8.00
	Sub-Total	$ 6,820.00

Present Practice

Set-up Cost/40 Subjects		$0
Maintenance Cost/40 Subjects		0
Measurement - SATCOMP/40 Subjects		12.00
ITBSCONC/40 Subjects		8.00
	Sub-Total	$ 20.00

Salary

Team #5 Salary/8 Weeks/2 Research Persons		$ 5,600.00
	Sub-Total	$ 5,600.00

Security

Security Deposit in Escrow		$ 2,000.00
	Sub-Total	$ 2,000.00

Miscellaneous

Contingency or Miscellaneous Expenses		$ 1,200.00
	Sub-Total	$ 1,200.00
	TOTAL	$27,660.00

APPENDIX 4A

Name__________________________

FINAL EXAM. ED. C655, L.S. Collet, Fall 1972.
Part 1. To be done in class.

1. Suppose you have two groups of scores which you wish to test for a significant difference in means. If there is an equal number of scores in the two groups, it is always possible to arrange the scores in pairs. Thus, it is possible to calculate either a t test for independent samples or a t test for dependent (paired) samples. Use the example at left to show how you would calculate each type of test. It is not necessary to complete all the computations -- just show the formula you would use in each case, and substitute the correct numbers from the data given at left.

Group 1 X	Group 2 Y	Difference (X-Y)
6	3	3
5	4	1
7	5	2
9	6	3
8	7	1
$\Sigma X = 35$	$\Sigma Y = 25$	$\Sigma D = 10$
$\Sigma X^2 = 225$	$\Sigma Y^2 = 135$	$\Sigma D^2 = 24$
$\Sigma x^2 = 10$	$\Sigma y^2 = 10$	$\Sigma d^2 = 4$

Independent Samples.

Correlated samples (paired scores)

---

2. Explain how you can tell whether to use a correlated (matched pair) or independent t test, and give an example of a situation in which each test is appropriate.

---

3. It is also possible to compute the significance of the difference in group means using an analysis of variance (one way classification). Use the Winer computation formulas at left to complete the summary of analysis table on the following page. Then use the results to answer questions 4 and 5.

Winer Formula

(1) $G^2/pn + = 60{*}60/(2{*}5) = \underline{360}$

(2) $\Sigma X^2 = 3{*}3+4{*}4+\ldots+8{*}8 = \underline{390}$

(3) $(\Sigma A^2)/n \; \frac{25^2+35^2}{5} = \underline{370}$

SUMMARY OF ANALYSIS OF VARIANCE

Source	SS	df	MS	F
Total	______	__	______	*****
A:between	______	__	______	
Error	______	__	______	

---

4. Compare and contrast the F ratio and the t ratio as tests of the significance of differences in group means.

5. Using the data from examples given in Items 1 and 3, explain how to test for homogeneity of variance in the two groups. (Use F_{max})

6. Explain the purpose behind a test for homogeneity of variance, and the procedure which should follow if the test is: (a) significant (b) not significant.

7. Explain the difference between accepting the null hypothesis and failing to reject the null hypothesis. Which terminology do you recommend, and why?

8. What is the meaning of "random selection" and why is it important to experimental design?

9. Explain the meaning of:"critical value of t (or z)","critical region", and "region of rejection". (Use a diagram).

10. Name four levels of measurement and explain why it is important to be able to distinguish among them.

11. Explain the difference between independent variables and dependent variables, and give an example to show how these terms apply to a research problem.

12. Compare and contrast type 1 error and type 2 error.

PART TWO, FINAL EXAM, ED. C655. THIS SECTION DOES COUNT ON THE FINAL GRADE FOR THIS COURSE. CAREFULLY READ THE EXPERIMENT DESCRIBED BELOW, THEN ANSWER EACH OF THE QUESTIONS WITH REFERENCE TO THE INFORMATION GIVEN.

An educator had spent several years developing what he referred to as an Inquiry approach to teaching social studies. He has now developed a systematic method of translating a set of educational objectives into an "inquiry" lesson. He hypothesizes that any teacher who uses the method and follows the lesson plan developed will produce greater academic achievement in his students.

To test this hypothesis, the educator arranged a field experiment in a local school system. Twenty teachers, each of whom has a single social studies class at the junior high level, agreed to participate in the experiment. These teachers were randomly divided into two groups of ten teachers each. The first group was assigned to the experimental condition (inquiry lessons via the educator's programmed materials), and the second group were assigned to a control condition (lessons developed by whatever methods the teacher usually used).

The educator and the twenty teachers next selected a body of social-studies content which was appropriate for the grades concerned, but was not presently being taught by any of the teachers. Next, they cooperatively developed a set of instructional objectives based on the content, and two parallel forms of an achievement test measuring the extent to which these objectives were attained. A trial run with the tests, using students from another system which was studying the content in question, showed that the two test forms were highly correlated (r+ = +.93) and of equivalent difficulty (the mean score on test form 1 was 51.2 and on form 2 was 51.4).

Next, each teacher developed his lesson plan according to the condition to which he was assigned -- experimental teachers used the educator's lesson-development materials while control teachers used their usual method. The teachers all taught the chosen content by the lesson plans prepared (above) during the next eight weeks. At the beginning of the period, each student was administered form 1 of the achievement test, and at the end of the eight weeks each student wrote form 2 of the test. The gain score for each student was then calculated by subtracting the pre-test score from the post-test score.

ANSWER THE FOLLOWING QUESTIONS:

13. One of the educator's statistical advisors suggested that the gain scores for the two groups of classrooms be entered into one giant t test, with each group containing 10m students, where n is the number of students per class. (Assume, for purposes of this discussion, that each class had exactly n=30 students.) The second statistical advisor agreed that a t test could be used to compare gain scores in the two groups, but he recommended that the pooled estimate of sigma be based on the original 20 classrooms rather than on two big groups of 300 students each.

    1a. How many degrees of freedom would be associated with each t test?

    Method 1 ________df ; Method 2 ________df

    1b. Which method would result in the larger error, and why?

    1c. Which method would you recommend, and why?

14. Still another advisor recommended that the experiment be analyzed as a simple t test of the mean achievement gain for each teacher. He reasoned that entire classrooms were the appropriate unit of observation whenever the experimenter wished to generalize to a teaching method as opposed to a learning method. Thus in this case, the appropriate n was 10 in the experimental condition and 10 in the control condition. Obviously, the single best number to represent a whole class (or teacher) would thus be the mean of the gain scores for that class.

    2a. In the space below, sketch out a diagrammatic representation of this design, and show exactly how you would go about computing the t ratio. Use appropriate symbols and formulas throughout.

15. What is your critical reaction to the suggestion in 2, above? Explain fully.

APPENDIX 4B

GUIDELINES FOR JUDGING RESEARCH/EVALUATION PROPOSALS AND FINAL REPORTS USING THE FEHR-PRACTICUM RATING SHEET

The FEHR-PRACTICUM Rating Sheet is an analytic approach to judging the quality of a proposal or report. The overall strategy is to improve reliability and validity by requiring the rater to make separate judgments regarding the presence or absence of various characteristics considered by experts to be typical of high quality products. These characteristics are grouped according to the usual organizational components (e.g., introduction), and assigned arbitrary weights reflecting their relative importance to the overall rating.

Rating Instructions. Fill in the blank opposite each characteristic listed on your Rating Sheet using the scoring guidelines listed below. For each characteristic, use the rating opposite the statement which best describes the product being rated with respect to the characteristic concerned. When all the characteristics within a component have been rated, sum the characteristic ratings to obtain an overall rating, and place it in the box provided. IF THE OBTAINED SUM IS NEGATIVE, ASSIGN A RATING OF ZERO.

A. Preliminary Materials

1. Title page characteristics.

   (a) The problem is precisely identified in the title. Give:

   2 points if the title identifies the target population, the key dependent variable(s), and the critical comparisons to be made. Subtract one point for each of these elements to a minimum rating of zero.

   (b) The title is sufficiently concise for indexing. Give:

   1 point if the title has 20 words or fewer and contains at least three keywords which accurately reflect the contents; otherwise 0 points.

   (c) The title is too long or wordy. Give penalty of:

   -1 point if the title exceeds 25 words in length; no deduction otherwise.

   (d) The format of the title page is inappropriate and/or there is incomplete author information. Give penalty of:

   -1 point if the format does not conform to the prescribed standard or, in the absence of a prescribed standard, if the author is not identified; no deduction otherwise.

2. Tables of contents, figures, etc. Give:

   4 points if there is a complete table of contents listing every major heading in the text and listings which itemize all the figures and tables in the text.

3 points if both types of listings (above) are present but one is incomplete.

2 points if there is a complete table of contents but no listing of figures/tables OR if both contents and figures/tables are listed but both are incomplete.

1 point for an incomplete table of contents OR incomplete figures/tables listing.

0 points otherwise.

3. Characteristics of abstract. (Award all zeroes if length exceeds assigned maximum.)

(a) The study purpose is outlined in the abstract. Give:

2 points if it summarizes the major questions to be studied in terms of relationships among variables.

1 point if it summarizes the questions, but the explicit relations to be studied are unclear.

0 points otherwise.

(b) The target population is identified. Give:

1 point if the population to whom the results are generalized is identified.

(c) Major dependent variables identified. Give:

1 point if the number and type of students are described.

0 points otherwise.

(d) The design is outlined. Give:

2 points if the design is clearly and accurately synopsized.

1 point if a summary statement of design exists, but any one of the following is missing: sampling procedures, dependent variable(s), independent or moderator variables.

0 points if there is no attempt to describe the design OR if two or more of the above elements are missing.

(e) The analytic procedures are outlined. Give:

2 points if the statistical (or other analytic) procedure used is clearly identified.

1 point if the procedure is mentioned but it is unclear what was done.

0 points otherwise.

(f) The key comparisons are outlined. Give:

2 points if the critical contrasts are explicitly identified. (It is not necessary that they be labelled.)

1 point if the key contrasts are implied but not explicitly mentioned.

0 points otherwise.

B. Body of the Proposal

1. Characteristics of the introduction.

An introductory section need not be labelled, but it must appear within the first third of the body of the proposal or report. It would normally contain the following elements in any order and under any label or heading.

(a) A statement of the problem. Many writers have used the "statement of the problem" label as if it is synonymous to "background of the study" and/or "purpose of the study". However, in this document the three terms have distinct and rather unique meanings, as explicated by the scoring guides for items (a) through (c). It is important that the user rate these items along the delineated dimensions. Give:

4 points if there is an explicit statement of the "basic" or "root" problem. To rate full credit, the statement should identify, at least in general terms, each of the following:

(i) the system being studied
(ii) what is presently happening in the system
(iii) what should be happening in the system
(iv) the reason for believing that it should happen

Assign one point for each of the above elements present.

(b) A description of the context or background of the study. Give:

4 points if the questions: "Why was this study proposed?" and "What has been done in this area by previous workers?" are explicitly answered.

2 points if only one of the above questions is answered OR if the answers are implicit rather than explicit.

0 points otherwise.

(c) The purpose of the study is defined within the first third of the text. Give:

4 points if there is an explicit statement of the specific questions to be answered by the study AND all questions are stated in terms of relationships among variables AND the questions are consistent with the remaining text.

3 points if the questions are consistent with the text and stated explicitly but not as relationships among variables.

2 points if the questions are explicit but inconsistent with the text OR if the questions are consistent but stated implicitly rather than explicitly.

1 point if there is a section labelled "purpose", "problem statement", or some synonymous term, which states the questions to be answered, but most of the questions are vague or ambiguous.

0 points if more than one third of the document must be read to determine its purpose.

(d) The importance of the study is established. Give:

2 points if there is an explicit statement of the potential benefits of the study.

1 point if the statement is implicit rather than explicit.

0 points otherwise.

(e) The scope of the study is delimited. Give:

2 points if there is an explicit statement explaining why the study was focused on the particular population and variables chosen.

1 point if it is clear why these were chosen but no explicit explanation is made.

0 points otherwise.

(f) The major assumptions and limitations are identified. Give:

4 points if the introductory section contains an explicit mention of all the important assumptions which underlie the study AND/OR the important limitations and weaknesses of the study.

3 points if the above statement appears after the introductory section -- e.g. in the results or discussion sections.

2 points if there is implicit rather than explicit discussion of the assumptions and/or limitations OR if no discussion exists but the rater cannot identify potentially dangerous assumptions or limitations.

1 point if the rater can identify one critical assumption or limitation (i.e., one which would definitely change the thrust or interpretation or validity of the study) which has not been discussed.

0 points if there is no discussion of assumptions or limitations in the entire study AND the rater can identify important assumptions or limitations (i.e., ones that might change the thrust or interpretation or validity of the study).

NOTE. In many studies a section of text which provides definitions of terms with unique or restricted technical meanings appears in or near the introduction. Since these meanings are closely related to the overall strategy or conceptual framework of the study, they are evaluated under that heading rather than here.

(g) The material within the introductory section lacks logical interrelations. Give penalties of:

0 points if the material presented is smoothly connected and many of the above characteristics are present and individually meaningful, but there are inconsistencies, contradictions or ambiguities among characteristics.

-2 points if it would be necessary for the average member of the intended audience to read the section several times to determine what the study is about. (Do not impose this penalty if the re-reading is necessary because the reader does not have the background knowledge common to the writer's intended audience!)

-4 points if even after successive readings the average member would be uncertain about the study's purposes.

-8 points if after successive readings the average reader in the intended audience would have no idea what the study is about.

2. Characteristics of the review of the related literature.

(a) The articles reviewed are clearly related to the study. Give:

4 points if there are more than five articles (or reports or books) reviewed and all of them are clearly related to the study. When fewer than five studies are reported, full credit is given only if every study is at least marginally related AND there is evidence of a thorough search (e.g., Education Index, Psychological Abstracts, and ERIC for at least the last five years).

3 points if only 3-5 clearly-related articles are reviewed without evidence of a thorough search OR if there are more than five articles reviewed with the majority clearly related to the study and none absolutely irrelevant OR if 1-3 related articles have been reviewed but there is evidence that a thorough search has been made. (It is explicitly assumed that there will always be some relevant theory or practical experience to discuss.)

2 points if there are at least five clearly-related studies but also one or more absolutely irrelevant studies included OR if one or more absolutely irrelevant studies are included with fewer than five clearly-related studies and evidence of a thorough search.

1 point if only marginally related articles are presented without evidence of a thorough search.

0 points if no material is reviewed OR if none of the above statements apply.

(b) The methods (logical analyses, research procedures, and data analyses techniques) used in the reviewed articles are critically evaluated. Give:

4 points if the review indicates the methods used in each study and makes explicit substantive evaluations of their adequacy. However, it is not necessary to make such comments about each article separately; it is, in fact, preferable to group studies with common themes and/or methods and evaluate them as a group.

3 points if the substantive evaluation (above) occurs but the minority of the criticisms are not supported in context (but do appear logical).

2 points if the evaluations occur but more than half are unsupported, OR if there are well supported evaluations for about half the articles and none for the others, OR all evaluations occur but many are picayune or unsubstantiated.

(c) The articles reviewed are representative of the domain studied. Give:

4 points if there is evidence that the reviewed materials cover all the major developments in theory, research, and practice during at least the last five years which have a direct bearing on the study. Light coverage of an area is permissible only if there is explicit evidence that little has been done.

2 points if there is one of the above areas missing without evidence that no work has been done in the area.

1 point if two of the above areas are missing without evidence that no work has been done (e.g., suppose only the research articles have been reviewed).

0 points otherwise.

(d) The articles reviewed are grouped in logical order. Give:

4 points if the reviewed articles are grouped by common themes and evaluated and/or interpreted by groups in a logical order. If only four or five studies are presented, full marks could be obtained only if there is evidence of a thorough search -- in this case it is only necessary for the articles to be discussed in a logical order.

2 points if there is some avoidable redundancy in article descriptions and/or evaluations but it does not add more than 20% to the time required for reading the review.

0 points if the redundancy adds more than 20% to the time required for reading the review.

(e) The review is summarized and synthesized. Give:

4 points if there is a summary presented which points out the areas of agreement and disagreement among articles within each area (theory, research and practice), and demonstrates how the material from each of the areas relates to the problem being investigated.

2 points if the above summary exists but there is no explicit statement of its relationship to the problem OR if adequate summaries of the various areas (theory, research and practice) are present but no attempt is made to interrelate them.

0 points if neither of the above statements is applicable.

(f) Relevant studies are missing from the review. Give penalties as outlined to a maximum of -8.

-1 point for each missing directly-related article (book, etc.) which was listed during the last five years in Education Index, Psychological Abstracts, ERIC, or any other reference commonly used by the audience concerned.

-2 points for each missing article (book, etc.) from any source commonly available to the intended audience which would substantively alter the study or its interpretation.

3. Characteristics of the conceptual framework or rationale.

(a) There exists a statement of the principles from which the study plan derives. Give:

4 points if the study contains a section which clearly explains why each of the specific variable relationships (specific hypotheses) to be evaluated in the study was chosen. The section need not have a separate heading, but labels such as "rationale", "conceptual framework", "strategy", and the like are common.

3 points if the above explanations exist, but do not appear in a single unit of text (e.g., there is a separate rationale for each hypothesis).

2 points if there is an explicit attempt to explain each choice but the reasons for one or more of the selections remain unclear OR if there is no explicit explanation but all choices are explained in context.

1 point if there is no explicit explanation and most, but not all, of the choices are explained in context.

0 points if none of the above statements is applicable.

(b) The principles in (a) are derived from the theory and research reviewed. Give:

4 points if there is an obvious relationship between the reviewed literature (or the review summary) and the

conceptual framework OR if an explicit statement explaining the relationship is provided. To obtain full marks here, section (a) must have obtained at least a 2 rating (i.e., a $\geq$ 2).

2 points if no conceptual framework (i.e., a < 2) between the reviewed literature (or the review summary) and the variable relationships (hypotheses) to be evaluated is either obvious or explicitly explained.

0 points if neither of the above statements is applicable.

(c) The principles from which the study plan was derived form a coherent unit. Give:

4 points if there are listed principles (i.e., a $\geq$ 2) which fit together naturally or are explicitly integrated and synthesized to form a coherent viewpoint. A set of principles are coherent if data providing direct support for the validity of one principle tends to be supportive of every other principle.

2 points if most of the principles are coherent (in the above sense) but some appear to be entirely discrete and independent OR if there is no explicit statement of the conceptual framework (i.e., a < 2) but the variable relationships (hypotheses) to be evaluated form a coherent set.

0 points if neither of the above statements is applicable.

(d) The principal criteria get at the main purpose or objectives of an educational enterprise while the modifying criteria get at the practical background factors (such as cost, convenience, and time involved) and/or the unintended consequences of the enterprise (e.g., parent hostility). Give:

4 points if both kinds of criteria are included and an explicit distinction is made as to their use in interpreting data. (The labels "principal" and "modifying" need not be used.)

3 points if both kinds are included and their use is clear, but the distinction is implicit rather than explicit.

2 points if both kinds of criteria are present, but it is not clear how they will be used in "solving" the stated problem.

0 points if none of the above statements is applicable.

(e) The substantive research hypotheses are stated, or, in the case of a non-experimental study, the probable result patterns[1] are stated and the implications of each pattern explained. Give:

4 points if there is a set of explicit and unambiguous statements of substantive hypotheses or probable result patterns which is consistent with the purpose(s) of the study and which provides comprehensive coverage of the questions the study was intended to answer. In addition, each hypothesis or result pattern should be:

- (i) referenced to a specific target population.
- (ii) stated in terms of relationships among variables.
- (iii) concerned with observable variable and/or operationally defined constructs.
- (iv) (hypotheses but not result patterns should be) stated in an "if ... then" form.

3 points if elements (i) and/or (iv) are missing OR if element (iii) is missing for a minority of variables.

2 points if any two of the following element-sets are missing: (i) and/or (iv), (ii), (iii); OR if there is a set of statements which possess all the characteristics of a 4 rating except that the set covers a majority but not all of the questions which the study was intended to answer.

1 point if there is a recognizable attempt to provide a statement of substantive hypotheses or result patterns which does not possess enough of the listed character-elements to merit a 2 rating.

0 points if none of the above statements are applicable.

(f) The specific or unique terms used in the study are defined. Give:

2 points if all the terms encountered should be clear to the intended audience because they are already familiar or because they have been defined (either in context or in a specially labelled section).

1 point if there is an explicit attempt to define terms, but it is incomplete or ambiguous.

---

[1] The term "probable result patterns" refers to the particular kinds of interrelationships among variables for which the experimenter intends to search. It is preferable for the educational meaning of each of these patterns to be pre-specified for the same reasons as planned comparisons are preferable to post hoc comparisons in an experimental study.

(g) There is an explicit criterion of success. Give:

4 points if there appears in the text a statement or statements which either explicitly or implicitly define(s) a decision rule for determining whether the purposes of the study have been fulfilled.

4. Characteristics of the method or procedure.

(a) The subjects are described. Give:

2 points if there is a description of the pool of subjects from which the research samples were chosen. It should specify the distribution of characteristics salient to the problem (usually such things as age, educational level and the like). If this pool of subjects is not the (entire) target population, per se, there must also be an assessment of its representativeness of that population.

1 point if there is a description, but it omits one or more salient characteristics.

0 points if neither of the above statements is applicable.

(b) The sampling procedure is described. Give:

2 points if the description which is sufficiently detailed to permit replication.

1 point if there is a clear description, but insufficient detail for replication.

0 points if neither of the above statements is applicable.

(c) The sampling is representative. Give:

4 points if the sampling will allow valid generalization to the target population OR if a rational argument for assuming valid generalization is presented.

2 points if there is a mild bias in the representativeness of the sampling, but this should not affect validity.

0 points if neither of the above statements is applicable.

(d) The design of the study is described. Give:

4 points if the design is described with sufficient detail and accuracy to permit complete replication.

3 points if there is sufficient description to permit replication of the main elements of the design but some details are missing.

2 points if there is a coherent design description but it would not permit replication of one or more of the main design elements.

1 point if there is a section labelled "design", but it is ambiguous or unclear.

0 points if none of the above statements is applicable.

(e) There is design rationale. Give:

2 points if there is a section which:

(i) explains why the particular design was chosen.
(ii) assesses the validity of the design chosen.

Subtract one point for each of the above elements missing.

(f) The variables are not operationally defined. Penalties are assessed for each dependent, independent, or moderator variable which is NOT operationally defined in terms of observable criteria. Give penalties of:

-2 points for each variable concerned with a primary hypothesis.

-1 point for each variable concerned with a secondary hypothesis.

THE MAXIMUM PENALTY IS -4.

(g) The design provides the critical comparison groups. Give:

2 points if the design provides for a control group and separable groups for each treatment to be assessed.

0 points if the above statement does not apply.

(h) The design provides for valid comparisons. Give:

2 points if all critical comparisons implied by the (delimited) objectives (problem or purpose) of the study can be assessed within the design. If there are possible confoundings, a rational argument for assuming the effects of confounded variables are negligible must be given.

1 point if confoundings occur without supporting arguments, but such arguments could be made.

0 points if neither of the above statements is applicable.

(i) Some sources of invalidity are uncontrolled. Give a penalty of:

-2 points for each uncontrolled source of invalidity which threatens the main purposes of the study.

-1 point for each uncontrolled source of invalidity which threatens the secondary purposes of the study.

THE MAXIMUM PENALTY IS -4.

(j) The instrumentation is described. Give:

2 points if each instrument (test, questionnaire, observation) is described.

1 point if most but not all instruments are described.

0 points if neither of the above statements is applicable.

(k) Instruments are assessed for reliability and validity. Give:

4 points if there is an explicit assessment of the reliability and validity of each instrument used.

3 points if there is only an assessment of validity (for one or more instruments).

2 points if there is only an assessment of reliability for one or more instruments, OR if there are complete assessments for a majority of instruments.

1 point if there is any explicit assessment of reliability or validity for one or more instruments.

0 points if none of the above statements is applicable.

(l) The instrumentation is unsuitable. Give a penalty of:

-2 points for each instance of an instrument which is invalid for its intended use.

-1 point for each instance of an inappropriate but not (completely) invalid use of an instrument.

0 points if neither of the above statements is applicable.

(m) The data collection procedures are described. Give:

2 points if the questions "which instruments?", "who administered?", "when", and "to whom" are answered for each data set.

1 point if any three of the above questions are answered.

0 points if neither of the above statements is applicable.

(n) The data matrix is defined. Give:

2 points if there is a schematic representation of the data matrix OR if the description is complete enough to permit such a schematic to be constructed.

1 point if there is an inaccurate or incomplete schematic.

0 points if neither of the above statements is applicable.

(o) The analytic procedure is described. Give:

4 points if the description is complete enough to permit replication of the analysis and if there is a rationale explaining why the procedure was considered most appropriate.

3 points if only the rationale is missing from the above but the procedure concerned is commonly used for similar purposes.

2 points if only the rationale is missing from the above and the procedure concerned is not commonly used, OR if a rationale is present but the description is insufficient to permit replication of the analysis.

1 point if there is an attempt at describing the analytic procedure which does not satisfy any of the above statements.

0 points if none of the above statements is applicable.

(p) The analysis evaluates all hypotheses. Give:

4 points if every hypotheses is explicitly evaluated by some contrast or measured relationship. (This need not be a valid contrast or measure to obtain marks.)

2 points if all primary hypotheses are directly evaluated but one or more secondary hypotheses are evaluated indirectly, OR if there are redundant (statistical) tests using a priori probabilities.

0 points if neither of the above statements is applicable.

(q) The analysis is efficient. Give:

4 points if the analysis uses the minimum valid estimate of error in evaluating comparisons. That is, it maximizes the statistical power of the test (without changing the significance level).

3 points if the analysis is the most efficient (powerful) of the procedures available to the researcher (e.g., univariate ANOVA when MANOVA is called for but not available on the local computer).

2 points if the analysis is not the most efficient (powerful) available, but it is reasonably efficient and/or consistent with common practice.

0 points if none of the above statements is applicable.

(r) The analytic procedures are inappropriate or invalid for the study's purpose(s). Give penalties of:

-2 points if the procedure will probably lead to an erroneous conclusion with respect to one secondary hypothesis.

-4 points if the procedure will probably lead to an erroneous conclusion with respect to more than one secondary hypothesis.

-6 points if the procedure will probably lead to an erroneous conclusion with respect to one important hypothesis but is sound with respect to the study's major purpose.

-8 points if the procedure will probably lead to an erroneous conclusion with respect to one or more of the study's major purpose, but can provide some valid conclusions.

-10 points if the procedure cannot lead to any valid conclusions and will probably lead to erroneous conclusions with respect to the study's major purposes.

NOTE: Sections 5 to 7 would normally appear in proposals but not in final reports.

5. Characteristics of the budget.

(a) The source of each item estimate is clear. Give:

2 points if it is obvious how each estimate was computed.

1 point if it is obvious for most items.

0 points if neither of the above statements is applicable.

(b) The standard items are present. Give:

2 points if all items in the guidelines given by the funding agency are present.

1 point if all items are covered but the itemization differs in unimportant ways from the guidelines.

0 points if the itemization differs substantively from the guidelines.

(c) Probable costs of delays or increased prices/wages are anticipated. Give:

2 points if the effects of inflation/deflation and probable delays.

1 point if an attempt has been made but it is incomplete.

0 points if there is no attempt or an inadequate attempt.

(d) The expenses and probable resources balance the needs of an adequate project. Give:

2 points if the budgeted amount appears reasonable for the purpose concerned.

1 point if the amount is too low to permit an adequate job or too high to be justified providing the deficiency or excess does not exceed 20% of the total contract.

0 points if the amount is deficient or excessive by factors greater than 20%.

(e) The cost effectiveness of the proposed study is assessed. Give:

2 points if there is an explicit and comprehensive attempt to demonstrate the cost effectiveness of the proposed project.

1 point if there is an explicit attempt which is less than comprehensive.

0 points if neither of the above statements is applicable.

6. Characteristics of logistics section.

   (a) A schedule of activities is provided. Give:

   2 points for a comprehensive schedule.

   1 point for a less than comprehensive schedule.

   0 points if neither of the above statements is applicable.

   (b) The planned work distribution is proportional to the man-hours available. Give:

   2 points if these elements appear balanced throughout.

   1 point if there is a mild increase or decrease in work with no change in resources.

   0 points if there is a sharp increase or decrease in work with no adjustments to staff.

   (c) There are sufficient personnel available. Give:

   2 points if there is evidence that persons with the needed skills will always be available at the times needed.

   1 point if there is some possibility that competent personnel will not be available.

   0 points if it is likely that competent personnel will not be available as needed.

   (d) Possible bottlenecks have been anticipated. Give:

   2 points if all probable bottlenecks are explicitly planned for (OR if no probable bottlenecks exist).

   1 point if an incomplete plan for handling bottlenecks is present.

   0 points if neither of the above statements is applicable.

   (e) The proposed sequence is logical and efficient. Give:

   2 points if the sequence makes optimum use of resources, and appears likely to work smoothly and well.

   1 point if there is a workable plan with less-than-optimum use of resources.

   0 points if neither of the above statements is applicable.

7. Characteristics of personnel.

   (a) The major personnel are named. Give:

   3 points for a complete list.

1 point for an incomplete list.

0 points if neither of the above statements is applicable.

(b) The responsibilities of all major personnel are defined. Give:

4 points for a comprehensive definition of responsibilities.

2 points for an incomplete list of responsibilities.

0 points if neither of the above statements is applicable.

(c) There is evidence of the competencies possessed by each of the major personnel. Give:

3 points if there is a complete (summary) vita for each major personnel member.

2 points if one of the vita's is sketchy or incomplete (but not missing).

1 point if one vita is missing, OR if more than one vita ia sketchy or incomplete.

0 points if none of the above statements is applicable.

(d) The (major) personnel are inadequate for the proposed project. Give a penalty of:

-2 points if a minority of the personnel appear competent, but lacking in experience.

-4 points if a majority of the personnel appear competent, but lacking in experience.

-6 points if a minority of the personnel are lacking in competence with respect to their assigned tasks.

-8 points if a majority (but not all) of the personnel are lacking in competence with respect to their assigned tasks.

-10 points if the entire set of personnel appear to be lacking in most of the prerequisite skills.

NOTE: Sections 8 and 9 apply to a final report but not to a proposal.

8. Characteristics of the results (statistical conclusions).

(a) There is a result presented for each hypothesis (or relation). Give:

4 points if every results are explicitly presented for each hypothesis.

3 points if all hypothesis results are covered, but some are implicit rather than explicit.

2 points if results are presented for all but a minor or secondary hypotheses.

1 point if most hypothesis results are presented, but some important hypothesis results are not.

0 points if none of the above statements is applicable.

(b) Explicit statistical conclusions are stated for each hypothesis. Give:

2 points if each result presented includes a statement (either in the text or in a table) of the significance or non-significance of the comparison or relationship evaluated and the direction of all significant findings.

1 point if significance, but not directionality, is presented for one or more of the results, OR if significance was presented implicitly, but not explicitly for some of the results.

0 points if neither of the above statements is applicable.

(c) There are neat concise displays of all results. Give:

2 points if all results are presented in neat, concise style with tables used whenever this was advantageous.

1 point if a minority of results were presented in unnecessarily redundant or wordy style OR if the results are complete but tables would have added to the clarity and/or conciseness.

0 points if neither of the above statements is applicable.

(d) The organization of the results is logical. Give:

6 points if the results are organized in clear, logical, easy-to-read style which minimizes the need for recursive reading (looking back).

4 points if the overall results are clear, but either the style requires recursive reading which adds less than 25% to the reading time required, or if there are minor ambiguities in the text caused by poor connectives or poor sequencing.

2 points if the organization and style requires an amount of recursive reading which increases reading time by 25-50%, or if major ambiguities are caused by poor connectives or poor sequencing.

0 points if none of the above statements is applicable.

(e) Explanatory graphs or diagrams are used to clarify meaning. Give:

4 points if all needed graphs (etc.) were included in clear readable form.

3 points if the needed graphs (etc.) are present but could be improved in format.

2 points if some needed graphs (etc.) are absent but those presented are clear and readable.

1 point if there is any use of graphs or diagrams which adds to clarify but does not satisfy any of the above statements.

0 points if none of the above statements is applicable.

(f) The results are summarized, conflicts are reconciled, and an overall synthesis provided. Give:

12 points if there are portions of text which clearly and succinctly summarize and synthesize all the results presented.

9 points if the summary is complete but the text could be more succinct, OR if there is a succinct summary and interpretation of individual findings but only a weak synthesis, OR if there is a succinct, complete summary, and an adequate synthesis, but a minor conflict in results has not been resolved.

6 points if all results are summarized and interpreted separately, but there is no attempt to interrelate or synthesize the findings.

3 points if there is a section labelled "summary" or the like which does not satisfy any of the above statements.

0 points if none of the above statements is applicable.

(g) There are procedural errors in the results -- that is errors which will always lead to erroneous conclusions. Give penalties of:

-2 points for each inaccurate or invalid statistical conclusion and/or each invalid interrelationship (synthesis) of statistical conclusions.

-5 points additional penalty if the major reported conclusion(s) are in error.

THE MAXIMUM PENALTY IS -10.

9. Characteristics of the educational conclusions and implications.

(a) An educational meaning is provided for each statistical conclusion. Give:

4 points if explicit educational interpretations are provided for all statistical conclusions (not necessarily separate or in the same order).

3 points if all interpretations are provided but one or more are implicit rather than explicit.

2 points if all but a few minor interpretations are provided.

1 point if an obvious attempt has been made to provide interpretations, but none of the above statements is satisfied.

0 points if none of the above statements is applicable.

(b) The discussion and presentation is objective, not subjective. Give:

4 points if the presentation is entirely objective, free from biases such as selection of only agreeable facts or treating all unconfirmed hypotheses as type 2 errors.

3 points if there are occurrences of subjectivity, but these do not substantively affect conclusions.

2 points if the presentation is objective except for the treatment of one or more unconfirmed hypotheses as if it were necessarily due to a type 2 error -- that is the non-significance was considered due to poor instrumentation, small n and the like without entertaining the possibility that the effects really were zero.

0 points if none of the above statements is applicable.

(c) The pattern of results is interpreted. Give:

4 points if there is an explicit and logical attempt to integrate the overall meaning of the pattern of results (as opposed to a discrete interpretation of each separate finding) that was "built in" to the design and analysis on an a priori basis.

3 points if there is an explicit and logical integration of the pattern of the results on a post hoc basis.

2 points if there is any explicit attempt to interpret the overall pattern as an entity which does not satisfy either of the above statements.

0 points if none of the above statements is applicable.

(d) The cost effectiveness of the various decision alternatives are assessed. Give:

8 points if there is an explicit assessment of the cost effectiveness which possesses the following features:

- (i) it includes all the important dependent variables available (must be more than one).
- (ii) it defines the relative importance of each dependent variable.
- (iii) it provides a rule or formula for transforming the raw multivariate data into a single score with interval properties which represents the degree to which the overall objectives has been met.

(iv) it provides comparable data on the cost of each decision alternative.

(v) features (iii) and (iv) are combined to give a single numeric representation of cost effectiveness.

6 points if there is an explicit assessment which possess all but features (iii) and (v) above, but which does possess a decision rule which permits all possible outcomes to be ordered (but not scaled as above).

4 points if there is an explicit decision rule which orders the obtained (but not all possible) outcomes according to criteria which involve both features (i) and (iv).

2 points for any explicit attempt to assess cost-effectiveness which does not satisfy any of the above statements.

0 points if none of the above statements is applicable.

(e) The conclusions and implications are valid. Give:

4 points if the conclusion and implication are complete, and each is valid for the populations specified (or if not specified, for the original target population).

3 points if the major conclusions are complete and valid, but there is some question about the validity of one or more secondary conclusions or the implications.

2 points if the conclusions and implications are valid, but some important and rather obvious conclusions/interpretations are omitted, OR if they do not apply to all members of the specified population.

1 point if there are explicit conclusions which have not been generalized beyond the experimental data, but which do not satisfy any of the above statements.

0 points if none of the above statements is applicable.

(f) There are misinterpretations of the results of the analysis. Give penalties of:

-1 point for each misinterpretation which does not affect the substantive conclusions.

-2 points for each misinterpretation which affects a substantive conclusion, but does not change the major conclusions, decisions or recommendations.

-5 points for each misinterpretation which affects the major conclusions, decisions or recommendations.

NOTE: Section 10 is an overall qualitative judgment which applies to both proposals and final reports.

10. Characteristics of the general evaluation.

(a) The study is physically neat and orderly. Give:

2 points if the entire study is neat and orderly.

1 point if most of the study is neat and orderly, OR if the study is uniformly moderately neat.

0 points if neither of the above statements is applicable.

(b) The style is acceptable to the audience for which it was intended. Give:

5 points if it meets the style requirement in all respects.

3 points if there are minor deviations from the style requirements, but these require only a moderate amount of editing.

1 point if there are substantive changes, major reorganizations or additions necessary to meet the style requirements.

0 points if none of the above statements is applicable.

(c) Appropriate citations are given in the text. Give:

3 points if citations are given whenever other persons work is used.

2 points if citations are given for major works, but not for those used for secondary purposes.

1 point if one (but not all) of the citations related to a major study purpose is omitted.

0 points if more than one citation related to a major purpose is omitted.

(d) The organization makes the study as a whole, clear and readable. Give:

5 points if it is of superiorclarity and readability.

4 points if it is of good clarity and readability.

3 points if it is of adequate clarity and readability.

2 points if it is of less-than-adequate clarity and readability.

1 point if it is of poor clarity and readability.

0 points if it is of unacceptable clarity and readability.

(e) The study as a whole is replicable. Give:

5 points if the entire study can be completely replicated.

4 points if the entire study can be replicated except for unimportant details.

3 points if the major themes can be replicated but the minor themes cannot.

2 points if most of the major themes can be replicated but some cannot.

1 point if at least one major theme is replicable but none of the above statements is applicable.

0 points if no major theme of the study is replicable.

C. Supplementary Materials

1. Bibliography. Rate the adequacy of the bibliography according to the following key:

   5 Superior

   3 Adequate

   1 Inferior

   0 No bibliography

2. Rate the additional explanatory powers of the appended data according to the following key:

   10 Compensates for most weaknesses in the text.

   8 Compensates for one major weakness or a majority of weaknesses in the text.

   6 Compensates for a number of important weaknesses in the text.

   4 Compensates for one important weakness in the text.

   2 Minimal explanatory power added.

   0 No explanatory power added.

APPENDIX 4C

Name____________________________

Self Assessment of Research and Evaluation Skills

QUESTIONNAIRE

THE FOLLOWING ITEMS SAMPLE A WIDE RANGE OF KNOWLEDGE AND SKILLS IN ORDER TO ASSESS THE DIFFERENCES BETWEEN THE TWO DIFFERENT LAB TECHNIQUES WE HAVE USED THIS TERM. THE ANSWERS WILL IN NO WAY AFFECT YOUR COURSE GRADE -- NOBODY IS EXPECTED TO BE COMPETENT IN ALL THE SKILLS TESTED. NEVERTHELESS, WE ASK THAT YOU DO YOUR BEST TO REPRESENT YOUR COMPETENCE, INTEREST, AND IMPORTANCE RATINGS ACCURATELY. THEY WILL BE OF GREAT HELP IN EVALUATING THE LABORATORY TECHNIQUES. ALL RESULTS WILL BE AVAILABLE TO INTERESTED STUDENTS EARLY IN JANUARY.

INSTRUCTIONS.

On page 2 are a list of tasks, with three blanks appearing before each listed task. Please indicate your competence, interest, and importance rating for each task by entering a number from 1 to 5 in each blank according to the following key:

Under COMP (blank 1) Indicate your competence to do the task listed by entering:

1 -if you have no competence, are completely unable to do the task.
2 -if you have minimum competence, can do the task with great study or by hiring a consultant.
3 -if you have moderate competence, can do the task acceptably with a minimum amount of study, can do it well with considerable study and little outside help.
4 -if you have high competence, can do the task well with minimal study, can do exceptionally well with extensive study and no outside help.
5 -if you have superior competence, can do an exceptionally fine job with only minimum (or no) study.

Under INT (blank 2) Indicate your interest in doing this sort of task by entering:

1 -if you have negative interest, find the task repugnant, wish to avoid it.
2 -if you have no interest in the area, but do not actively avoid it.
3 -if you find the area somewhat interesting, or desire to do it to reach some desirable end -- even though the task *per se* is not appealing to you.
4 -if you find the area moderately interesting *per se*, or highly desirable as a means to an end.

5 -if you are highly interested in the task in and of itself.

Under IMP (blank 3) Indicate the importance of this task in your career, as presently planned, by entering:

1 -if you believe the task is of no importance in your planned career, (or if you consider it irrelevant to your professional performance.
2 -if the task is of minimal importance, needs performed only occasionally, and/or the task is an unimportant aspect of your professional performance.
3 -if the task is moderately important, is relevant but not crucial to your professional performance.
4 -if the task is highly important either because it must be done frequently or because it is of necessary to adequate professional performance.
5 -if the task is crucial to adequate professional performance, regardless of frequency.

COMP	INT	IMP	LIST OF TASKS:
(_)	(_)	(_)	1. Choose between the independent and matched pair t test, and perform all computations.
(_)	(_)	(_)	2. Compute and interpret a one-way analysis of variance.
(_)	(_)	(_)	3. Compute and interpret a two-way analysis of variance.
(_)	(_)	(_)	4. Compute and interpret a correlation coefficient.
(_)	(_)	(_)	5. Use a "canned" computer program to do simple analyses such as 1-4, above.
(_)	(_)	(_)	6. Select a sample randomly, and/or use stratified random sampling.
(_)	(_)	(_)	7. Discuss the concept of regression towards the mean and it affects a given experiment of your own design.
(_)	(_)	(_)	8. Describe the central limit theorem and suggest its implications for an experiment of your own design.
(_)	(_)	(_)	9. Contrast statistical significance with substantive (educational) significance, and give an example of each.
(_)	(_)	(_)	10. Compare and contrast type 1 and type 2 error.
(_)	(_)	(_)	11. Select samples and analytic procedures to optimize the probabilities of type 1 and type 2 errors for a particular problem of interest.
(_)	(_)	(_)	12. Compare and contrast four levels of measurement, and classify any given example according to its level of measurement.
(_)	(_)	(_)	13. Compare and contrast null hypothesis, one-tailed hypothesis, and two-tailed hypothesis.

COMP	INT	IMP	LIST OF TASKS:
(_)	(_)	(_)	14. Select cut-off points on a continuous selection variable (such as IQ) so as to have a .05 probability of excluding a student whose true score was equal to or less than a given score (e.g. 85 IQ).
(_)	(_)	(_)	15. Conduct an N-way analysis using a canned computer program.
(_)	(_)	(_)	16. Conduct a one-way analysis of covariance using a canned computer program.
(_)	(_)	(_)	17. Prepare data for analysis by computer by recording it in a usable format on computer punch cards, or by recording it on a tape or disk file.
(_)	(_)	(_)	18. Describe the meaning of statistical power, and recommend 2 ways to increase power.
(_)	(_)	(_)	19. Compare and interrelate confidence intervals, critical region, region of rejection and level of significance.
(_)	(_)	(_)	20. Describe the effect of truncation on a correlation coefficient and find some articles or reported research in which truncation (e.g., a ceiling effect or floor effect) have caused errors of interpretation.
(_)	(_)	(_)	21. Compare and contrast standard error of measurement, standard error of the mean, and standard error of the difference (in means).
(_)	(_)	(_)	22. Describe the relationship between reliability and validity.
(_)	(_)	(_)	23. Describe the concept of degrees of freedom, and give a rule for finding the df in a given case.
(_)	(_)	(_)	24. Distinguish among dependent variables, independent variables, predictors, and criteria.
(_)	(_)	(_)	25. Distinguish among main effects, interactions, simple main effects, and confounded effects.
(_)	(_)	(_)	26. Write a researchable hypothesis in operational terms.
(_)	(_)	(_)	27. Decide from the nature of the question what the design should be -- even though you may not be competent to analyze it.
(_)	(_)	(_)	28. Distinguish between internal and external validity for an experiment, and give several sources (threats) of each type of validity.
(_)	(_)	(_)	29. Distinguish between ex-post-facto post hoc experiments and a-priori experiments, giving the advantages and disadvantages of each.
(_)	(_)	(_)	30. Contrast a linear and curvilinear relationship between variables, and give an example of each.
(_)	(_)	(_)	31. Compare and contrast norm-referenced and criterion-referenced tests and give examples of legitimate uses of each type of measure.

COMP	INT	IMP	LIST OF TASKS:
☐	☐	☐	32. Distinguish between obtrusive and unobtrusive measures, give an example of each, and name the advantages of each for experimental purposes.
☐	☐	☐	33. Obtain the necessary information to judge the worth and utility of a given test, as compared with the highest measurement standards.
☐	☐	☐	34. Design and conduct a study to establish the construct validity of a new test, i.e., to test the extent to which the pattern of results agree with the theory upon which the test is built.
☐	☐	☐	35. Critically review an experiment in accord with the highest standards of scholarship.
☐	☐	☐	36. Construct a questionnaire free of major sources of error.
☐	☐	☐	37. Organize and report a formal review of the literature in a field of interest to you.
☐	☐	☐	38. Prepare and submit a formal proposal for funded research to a government agency or educational foundation.
☐	☐	☐	39. Plan and execute an experiment to determine the extent to which a given educational program is meeting its goals.
☐	☐	☐	40. Complete a dissertation in my area of interest, according to the standards established by your program.

APPENDIX 4D

Education C655: LABORATORY EXERCISE #1

L. S. Collet

PART I: Data The ten students in Miss Smith's class were each given standardized texts of arithmetic (X) and geography (Y) with the results tabled below. Fill in the blanks below using the most efficient formulas and procedures for the data given.

	X	Dev. X	(Dev. X)^2	(Dev. X)(Dev. Y)	(Dev. Y)^2	Dev. Y	Y
Albert	48						69
Bernice	60						81
Cameron	66						81
Denise	60						93
Ellen	84						99
Frederick	66						93
Geneva	42						87
Harry	30						63
Ingrid	42						69
Janet	42						75
Sums							

Part II: Word Translation.

Statistic	Formula	Numbers	Answers
Mean of X			
Mean of Y			
Covariance X, Y			
Variance of X			
Variance of Y			
Standard Deviation X			
Standard Deviation Y			
Correlation X, Y			
Variance of (X-Y)			
Variance of 4Y			
Variance of (X-2)			
Mean of (X/3)			
Mean of (Y+7)			
Harry's Z score on Y			
Ellen's Z score on X			

Part III: Formula Translations.

Using the data from I, above, fill in the blanks with the number associated with each of the following identities.

$\frac{\Sigma X}{N}$ = ______

ΣY = ______

$N\bar{X}$ = ______

$N\bar{Y}$ = ______

$\frac{\Sigma X \Sigma Y}{N\bar{X}}$ = ______

$(N-1)S_X^2$ = ______

$\Sigma XY - \frac{\Sigma X \Sigma Y}{N}$ = ______

$\Sigma XY - N\overline{XY}$ = ______

$\frac{\Sigma xy}{(N-1)S_x S_y}$ = ______

$\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$ = ______

$\frac{\Sigma X^2}{N-1} - \frac{N(\bar{X})^2}{N-1}$ = ______

$\frac{\Sigma xy}{N-1}$ = ______

$\Sigma Z_x Z_y$ = ______

$(N-1)$ COVxy = ______

$\frac{\text{COVxy}}{\Sigma xy}$ = ______

$\frac{\Sigma Z_x Z_y}{N-1}$ = ______

$\Sigma X^2 - N\bar{X}^2$ = ______

$\sqrt{\frac{N\Sigma Y^2 \ (\Sigma Y)^2}{N(N-1)}}$ = ______

$$\frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}} = \underline{\qquad\qquad}$$

Part IV: Practical Application.

Suppose that the published norms for the two tests given in I above were:

Arithmetic: Mean = 50 S.D. = 6 N = 1000

Correlation . 75

Geography: Mean = 60 S.D. = 5 N = 2000

Answer the following questions:

(a) What was the variance of Arithmetic scores for the normative group?

(b) What was the sum of squared deviations of the mathematics scores for the normative group?

(c) What was the covariance of the arithmetic and georgraphy scores in the normative group?

(d) How many students in the normative group for arithmetic achieved scores equal to or less than the mean for Miss Smith's class?

(e) How many students in Miss Smith's class achieved scores exceeding a standard score of +1.0 in georgraphy for the normative group?

(f) Suppose that a frequency distribution had been prepared for the georgraphy scores of the normative group. What would be the sum of the frequency column?

NAME:_______________

C655

L. S. Collet

LABORATORY EXERCISE #2

Mr. Jones has developed a set of computer-assisted drill and practice lessons in arithmetic reasoning. In order to test the hypothesis that students can learn arithmetic reasoning from his lessons, he performs the following experiment with his math class. First, he constructed two 100 item tests of arithmetic reasoning by writing 100 pairs of equivalent items then randomly assigning one member of each pair to each test. At the beginning of the experimental period test 1 was administered (pre-test) and the scores recorded. During the next three weeks each student used three of his five math classes each week for computerized drill and practice lessons. The remaining math periods were spent on the regular program. At the end of the three weeks experimental period test 2 (post-test) was administered. The difference between post- and pre-test scores was considered to be the learning in arithmetic reasoning due to the drill. Answer each of the following questions:

(a) Calculate the mean, unbiased variance and standard deviation, and the standard error of the mean for pre, post, and gain scores.

(b) Calculate the Pearson product moment correlation of pre and post scores and the standard error of the difference in pre and post means.

(c) Compute the .95 confidence intervals for the means of pre, post, and gains scores.

(d) Test the significance of the difference in pre and post scores. State the null hypothesis, show the test, and state the statistical conclusions.

*(e) Write an educational conclusion.

Note. Show all steps. Engineers pad is preferred.

---

*Not a recorded item for this test. For practice purposes only.

LABORATORY EXERCISE #3

C655

NAME:____________________________

Title: An Empirical Comparison of Two Methods of Teaching The Descriptive Characteristics of Major Mental Disorders.

Problem: Does a lecture using visual aids produce more learning than a lecture without visual aids.

Procedure: The content to be taught was the descriptive characteristics of major mental disorders. First, a 30 minute silent film was prepared which visually illustrated the more obvious characteristics (panic, catatonia, and so on) of each of the disorders. Next, a lecture was prepared which served as a background commentary to the movie, but which in itself was completely meaningful without the movie.

The investigator was teaching two introductory classes in educational psychology. Since the classes were of comparable ability, he decided to use the regular lesson time to run his experiment. By flipping a coin he determined that the first class would get the movie and the second would not. On the Monday morning he read his lecture to class 1 with the movie running. Then he administered a 100 item multiple-choice test which required subjects to pick out the symptoms of a given mental disorder. On Tuesday he read the same lecture to class 2 without the movie running and again administered the test. To obtain your data, issue the following commands to the computer:

```
$RUN  K04A:SIMEX   4=K04A:FREQ    5=K04A:TEST1
000010(your soc. sec. #)00001
Your data will be printed out.
```

Do an appropriate statistical analysis and state your educational conclusions.

The data appears on the attached computer printout.

NOTE: This is due next week. Be sure to state your conclusions in good experimental form.

LABORATORY EXERCISE #4

```
THIS IS EXPERIMENT   1
ILLUSTRATIVE EXPERIMENT 1 , C655 TERM 1,1972. L.S. COLLET, INSTRUCTOR.
     JOHN AND MARY ARE PH.D. STUDENTS MAJORING IN REMEDIAL READING.
EACH HAS WRITTEN A PROGRAMMED TEXT DESIGNED AS A SELF-STUDY PROGRAM
IN REMEDIAL READING AT THE JUNIOR HIGH LEVEL. THEY DECIDE TO PERFORM
AN EXPERIMENT TO SEE WHICH WAS THE BETTER TEXT. THE SCHOOL ADMINIS
TRATION FURNISHED THEM WITH THE NAMES OF THE 160 STUDENTS IN THE
CITY JUNIOR HIGHS WHOSE GRADE SCORE ON THE ROUTINELY ADMINISTERED
GATES READING TEST WAS 5.0 OR LESS. THE 160 NAMES WERE PLACED IN A
HAT, THEN MARY DREW 80--LEAVING 80 FOR JOHN.
     JOHN DIVIDED HIS NAMES INTO TWO GROUPS BY CALLING THE FIRST
40 GROUP I AND THE SECOND 40 GROUP II. MARY DIVIDED HERS IN THE
SAME WAY TO OBTAIN GROUPS III AND IV.
     FOR THE NEXT MONTH, GROUPS I AND III STUDIED MARY'S TEXT AND
GROUPS II AND IV JOHN'S TEXT, WITH JOHN SUPERVISING I AND II,
AND MARY SUPERVISING III AND IV. ASSUME THAT ALL SUBJECTS ATTENDED
EACH SESSION. AT THE END OF THE EXPERIMENTAL PERIOD, A PARALLEL
FORM OF THE GATES READING TEST WAS ADMINISTERED. THE RESULTING
SCORES ARE TABULATED BELOW. COMPUTE THE APPROPRIATE T TESTS TO
ANSWER THE FOLLOWING QUESTIONS. IN EACH CASE TEST A DIFFERENCE BETWEEN
A PAIR OF GROUP MEANS. ALL SCORES BELOW ARE TABULATED AS WHOLE
MONTHS (TEN MONTHS = 1 GRADE)-- MOVE THE DECIMAL ONE PLACE LEFT
TO OBTAIN THE GRADE-SCORE. E.G., 55 EQUALS A GRADE SCORE OF 5.5 .
   1.WHICH TEXT WAS BETTER:  (A) UNDER JOHN'S SUPERVISION? (T TEST #1)
                             (B) UNDER MARY'S SUPERVISION? (T TEST #2)
                             (C) COMBINED OVER JOHN AND MARY?(T TEST #3)
   2. WHICH SUPERVISOR WAS MORE FACILITORY TO LEARNING? (T TEST #4)
   3. TEST THE INTERACTION, I.E., THE DIFFERENCE BETWEEN THE DIFFERENCES
      OBTAINED IN ITEM 1 PART A AND ITEM 1 PART B. (T TEST #5)
***** THE COLUMNS IDENTIFY GROUPS: C1=GRP 1,C2=GRP 2 ETC.

OBSERVED SCORES: OUTPUT BY GROUPS
            GROUP  1, ABC(111)
   56.     56.     58.     31.     25.     63.     49.     80.     66.     95.
   57.     39.     37.     60.     79.     34.     62.     52.     45.     57.
   76.     51.     52.     54.     53.     76.     52.     46.     45.     43.
   73.     62.     75.     51.     61.     79.     85.     75.     25.     50.
SUM=   2285.; SUM X2=   140943.; N= 40
            GROUP  2, ABC(121)
   31.     76.     81.     78.     95.     33.     55.     74.     84.     65.
   85.     98.     69.     78.     71.     84.     82.     93.     80.     79.
   86.     74.     73.     82.     81.     97.     64.     55.     54.     43.
   59.     85.     61.     47.     89.     48.     68.     94.    110.     83.
SUM=   2944.; SUM X2=   229154.; N= 40
            GROUP  3, ABC(211)
   54.     55.    110.     54.     70.     79.     60.     73.     47.     52.
   61.     70.     58.     87.     60.     77.     67.    110.     84.    106.
   68.    105.     66.     33.     77.     56.     94.     81.     88.     45.
   67.     33.     61.     49.     91.     86.    107.     98.    100.     27.
SUM=   2866.; SUM X2=   224388.; N= 40
            GROUP  4, ABC(221)
   79.     91.     58.     71.     69.     87.     60.    109.     67.     48.
   61.     81.     44.     54.     54.     57.     76.    103.     95.     74.
   36.     63.     41.     58.     67.     43.     92.     28.     52.     59.
   92.     52.     69.     56.     10.     35.     76.     66.     82.     48.
SUM=   2613.; SUM X2=   187347.; N= 40
```